FUNCTIONAL CONSTRAINTS AND *rbc*L EVIDENCE FOR LAND PLANT PHYLOGENY[1]

Victor A. Albert,[2,3] *Anders Backlund,*[2] *Kåre Bremer,*[2] *Mark W. Chase,*[4] *James R. Manhart,*[5] *Brent D. Mishler,*[6] *and Kevin C. Nixon*[7]

ABSTRACT

Although the proportion of "functional" DNA in eukaryotic genomes is both debatable and subject to definition, most sequences gathered for phylogenetic purposes are indisputably functional. For example, patterns of variation are likely to be strongly constrained in ribosomal RNAs because of their structural and catalytic roles in protein translation, and in protein-coding genes, because of protein function itself. Although seemingly obvious, these concerns are usually ignored by workers producing gene trees. We have examined the extent of functional constraints in land-plant *rbc*L sequences. Not only do *rbc*L sequences appear to change with essentially clocklike regularity, but nucleotide-based cladograms imply that approximately 97.5% of codon changes on internal branches are functionally neutral (i.e., synonymous or functionally labile). From this perspective, *rbc*L evolution appears to be strongly constrained by function. Transforming nucleotide data into ad hoc string recognitions alters the size of the unit character sufficiently to highlight "blocks" of conservative information that may or may not be functionally constrained. Simultaneous cladistic analysis of all available evidence will highlight the proportion of congruent information, despite diverse functional constraints among the characters analyzed. We demonstrate the strength of this approach using different forms of the same *rbc*L evidence (i.e., nucleotides, strings, or amino acids) in combination with the seed-plant data of Nixon et al.

Diversification of the major clades of extant land plants probably dates from the Silurian to Cretaceous. During the Silurian–Devonian, liverworts, hornworts, mosses, and tracheophytes formed distinct lineages. Differentiation of the tracheophyte clades, notably angiosperms and other seed plants, began by the Devonian. The estimation of land-plant phylogeny, a research goal spanning over 400 million years of cladogenesis and extinction, is no simple task. For example, many groups lack strong morphological similarities that might suggest patterns of relationship.

Recent years have seen an explosion of interest in molecular information, with its promise of easily interpreted similarities for bridging otherwise large phenotypic gaps. In particular, the plastid *rbc*L gene (which encodes the large subunit of RuBisCO: ribulose-1,5-bisphosphate carboxylase/oxygenase, a primary enzyme in carbon fixation) has been sequenced extensively, with primary emphasis on the angiosperms (Clegg, 1993; Chase et al., 1993). Arguing from expected synonymous substitutions per site under a particular rate assumption, Clegg (1993) suggested that *rbc*L sequences should be phylogenetically informative for the time interval 400–100 million years before present. We argue here that this and similar assertions are incomplete. From direct estimation of *total* substitutions (as optimized on cladograms; see Albert et al., 1992a, 1993; Albert & Mishler, 1992 Albert et al., 1993)

[1] We thank Steve Farris and Peter Engström for ideas, advice, and discussion, Diana Lipscomb, Brunella Bowditch, and Mick Richardson for comments on an earlier draft of the manuscript, Bill Crepet, Else Marie Friis, and Dennis Stevenson for permission to use unpublished versions of their data matrix produced in collaboration with KCN, and Bill Anderson for last-minute discussion of Malpighiaceae biogeography. All interpretations are, of course, our own. The program for constructing random strings was written by Karl-König Königsson and Rolf Staflin. Support from the Swedish Natural Science Research Council (to VAA, AB, and KB) and the U.S. National Science Foundation (BSR-8906496 to MWC, BSR-8906126 to JRM, BSR-9107484 to BDM) is gratefully acknowledged. Lastly, VAA thanks the symposium organizers, travel arrangement staff, and editorial office of the Missouri Botanical Garden for their courteous and patient assistance.

[2] Department of Systematic Botany, Uppsala University, Villavägen 6, S-752 36 Uppsala, Sweden.

[3] Present and corresponding address: Department of Physiological Botany, Uppsala University, Villavägen 6, S-752 36 Uppsala, Sweden.

[4] Laboratory of Molecular Systematics, Royal Botanical Gardens, Kew, Richmond, Surrey TW9 3AB, United Kingdom.

[5] Department of Biology, Texas A&M University, College Station, Texas 77843, U.S.A.

[6] University and Jepson Herbaria and Department of Integrative Biology, University of California, Berkeley, California 94720, U.S.A.

[7] L. H. Bailey Hortorium, Cornell University, Ithaca, New York 14853, U.S.A.

ANN. MISSOURI BOT. GARD. 81: 534–567. 1994.

we will demonstrate that divergence-time asymmetries among taxa restrict *rbc*L-based hypotheses of land-plant phylogeny far more than do rate asymmetries.

We have examined the internal stability of land-plant *rbc*L evidence through conversion of nucleotide information into different data forms, including presence/absence of ad hoc nucleotide strings. Cladograms produced from nucleotide, string, and translated amino acid data are only partially congruent. Character optimization on both nucleotide and string trees reveals extensive functional conservation through the predominance of silent changes and labile (function-conserving) amino acid replacements. Hence, *rbc*L nucleotides are no less functionally constrained than morphological characters (contra Olmstead, 1989; Sytsma et al., 1991; Clegg, 1993).

Although the separation of protein-functional from cladogenetic history may not be entirely possible, the extent to which functional history reflects phylogeny might be assessed through congruence studies with characters expected to carry diverse patterns of functional constraints. As such, we have performed total-evidence analyses at the seed-plant level using, as a "constant," a new matrix of primarily morphological data (Nixon et al., 1994, this issue). It emerges that combination of *rbc*L nucleotide, amino acid, or string data with this matrix produces highly compatible cladistic hypotheses. These studies point to (i) the commonality of information in different data forms representing the same evidence, and (ii) the power of simultaneous evaluation of all available evidence and weakness of further production of *rbc*L gene trees (cf. Kluge, 1989; Barrett et al., 1991; Donoghue & Sanderson, 1992; Jones et al., 1993; Mishler, 1994).

The Rate "Problem"

As has been pointed out in several recent papers, sequence change in the *rbc*L gene is not strictly clocklike (Albert et al., 1992a; Bousquet et al., 1992; Gaut et al., 1992; Clegg, 1993). Here, we provide a number of new comparisons (Table 1) based on patristic distances between woody taxon pairs from Search II of Chase et al. (1993). It is clear that our own estimates and those of other workers all fall within a very narrow range of absolutely low values. The mean rate per taxon pair investigated here is approximately 2×10^{-10} total substitutions per site per million years; Wendel & Albert (1992) estimated $5–7 \times 10^{-10}$ for three herbaceous-pair comparisons. Lineage-specific rate differences were found by Bousquet et al. (1992) and in the relative-rate tests of Gaut et al. (1992), but absolute rate estimates do not differ substantially from our own findings. Thus, whereas *rbc*L data cannot be considered perfectly ultrametric (i.e., satisfying a clock assumption), the small range of absolute variation suggests that some predictions of the clock hypothesis still apply. For example, the relationship between time and the accumulation of nucleotide substitutions may be nearly linear. We term this condition, apparently characterizing *rbc*L sequence data, "quasi-ultrametric."

Quasi-ultrametricity has several important implications. One is that the extent of sequence divergence in a given taxon sampling should roughly reflect the timing of underlying cladogenetic events. If all such events are ancient, extensive sequence differences among all taxa are to be expected (Fig. 1; cf. Donoghue & Sanderson, 1992, fig. 15.3). If some cladogenetic events are ancient whereas others are much more recent, expected sequence divergence in a data set would be prominently skewed (Fig. 2). As these properties become extreme, parsimony analysis will be hampered by the increased probability of parallel changes among either anciently diverged or divergence-time-asymmetric sequences (Figs. 1, 2; cf. Donoghue & Sanderson, 1992: 347–349). Given that A, T, G, and C are the only character-state alternatives, either scenario is likely to produce patterns of similarity that may be nonhomologous and therefore cladograms that are ahistorical. This is precisely the "long branches attract" issue raised by Felsenstein (1978) and others.

Although asymmetrical rates of sequence change are often invoked to explain branch attraction behavior (see Clegg and Zurawski, 1992: 10, with reference to *rbc*L), the problem is better defined in terms of both rate and divergence time as their product, per-character change: the λ of Albert et al. (1992a, 1993; Albert & Mishler, 1992; cf. Hendy & Penny, 1989). With quasi-ultrametric data, rate asymmetry is unimportant in this regard; time through which a branch exists becomes the central factor. As such, our expectation of the performance of parsimony analysis on *rbc*L data must include our ability to estimate both the absolute and relative timing of cladogenetic events inherent to particular data matrices. Of course, this may not always be possible.

An additional implication of quasi-ultrametricity is the near satisfaction of selective neutrality. A molecular clock is predicted by the neutral theory of molecular evolution; equal rates of mutation and fixation are the expectation (see Kimura, 1983;

TABLE 1. "Phylogenetic" estimation of total substitution rate for 19 woody-taxon pairs. The rate of sequence divergence was calculated as per-site divergence (the patristic distance, D_p, divided by the number of nucleotides compared) divided by time since cladogenesis (Albert et al., 1992a). Average rates for individual taxa are half of the values shown. Data are from Search II of Chase et al. (1993); systematic error associated with that analysis can be expected to affect all calculations equally. Divergence time assumptions are based upon geologic dates associated with vicariant disjunctions (with the exception of all Arecaceae comparisons, which follow from the arguments of Wilson et al., 1990).

Taxon pair	Area	Divergence time assumption	D_p	Divergence rate (subst./site· taxon pair)
Callitris rhomboidea R. Br. ex Rich. *Widdringtonia cedarbergensis* Marsh (Cupressaceae)	Australia Africa	100 My[a]	55	3.85×10^{-10}
Metasequoia glyptostroboides Hu & W. C. Chang *Sequoiadendron giganteum* (Lindl.) J. Buchholz (Taxodiaceae)	Asia N. America	40 My[b]	16	2.80×10^{-10}
Illicium parviflorum Michx. ex Vent *Austrobaileya scandens* C. T. White (Illiciaceae/Austrobaileyaceae)	N. America/Asia Australia	200 My[c]	54	1.89×10^{-10}
Drimys winteri J. R. & G. Forst. *Belliolum* sp. (Winteraceae)	S. America New Caledonia	100 My	21	1.47×10^{-10}
Drimys winteri J. R. & G. Forst. *Tasmannia insipida* DC. (Winteraceae)	S. America Tasmania	100 My	14	0.98×10^{-10}
Canella winteriana (L.) Gaertn. *Belliolum* sp. (Canellaceae/Winteraceae)	N. America New Caledonia	200 My	78	2.73×10^{-10}
Canella winteriana (L.) Gaertn. *Tasmannia insipida* DC. (Canellaceae/Winteraceae)	N. America Tasmania	200 My	67	2.35×10^{-10}
Liriodendron tulipifera L. *Liriodendron chinense* (Hemsl.) Sarg. (Magnoliaceae)	N. America Asia	40 My	10	1.75×10^{-10}
Calycanthus chinensis Cheng & S. T. Chang *Idiospermum australiense* (Diels) S. T. Blake (Calycanthaceae/Idiospermaceae)	Asia/N. America Australia	200 My	28	0.98×10^{-10}
Chimonanthus praecox (L.) Link *Idiospermum australiense* (Diels) S. T. Blake (Calycanthaceae/Idiospermaceae)	Asia Australia	200 My	24	0.84×10^{-10}
Chamaedorea costaricana Oerst. *Drymophloeus subdistichus* (H. E. Moore) H. E. Moore (Arecaceae)	Americas S. Pacific	60 My[d]	15	1.75×10^{-10}
Chamaedorea costaricana Oerst. *Nypa fruticans* Wurb. (Arecaceae)	Americas S. Pacific/India	60 My	20	2.33×10^{-10}
Serenoa repens (Bartram) Small *Drymophloeus subdistichus* (H. E. Moore) H. E. Moore (Arecaceae)	Americas S. Pacific	60 My	18	2.10×10^{-10}

TABLE 1. Continued.

Taxon pair	Area	Divergence time assumption	D_p	Divergence rate (subst./site· taxon pair)
Serenoa repens (Bartram) Small *Nypa fruticans* Wurb. (Arecaceae)	Americas S. Pacific/India	60 My	23	2.68×10^{-10}
Betula nigra L. *Casuarina litorea* L. (Betulaceae/Casuarinaceae)	N. Hemisphere Australia	200 My	35	1.23×10^{-10}
Nothofagus dombeyi (Mirb.) Oerst. *Nothofagus balansae* (Baill.) Steenis (Nothofagaceae)	S. America New Caledonia	100 My	30	2.10×10^{-10}
Galphimia gracilis Bartl. *Acridocarpus natalitius* A. Juss. (Malpighiaceae)	S.-N. America[e] Africa/Madagascar/ New Caledonia	100 My	34	2.38×10^{-10}
Dicella nucifera Chodat *Acridocarpus natalitius* A. Juss. (Malpighiaceae)	S. America Africa/Madagascar/ New Caledonia	100 My	33	2.31×10^{-10}
Mascagnia stannea (Griseb.) Nied. *Acridocarpus natalitius* A. Juss. (Malpighiaceae)	S.-N. America Africa/Madagascar/ New Caledonia	100 My	34	2.38×10^{-10}
Range				3.01×10^{-10}
Mean				2.05×10^{-10}
S.D.				$\pm 0.75 \times 10^{-10}$

[a] Standard time figure used to represent the breakup of Gondwana (rounded to the nearest 100 My (million years) from 130 My, as estimated using Terra Mobilis® 2.1 by C. R. Denham and C. R. Scotese; see Wendel & Albert, 1992: 137).

[b] Standard time figure (ca. early Oligocene) used to represent disruption of the boreotropical interchange between North America and Eurasia (see Lavin & Luckow, 1993).

[c] Standard time figure used to represent separation of the Northern and Southern Hemispheres upon the breakup of Pangaea (rounded to the nearest 100 My from 160 My, as estimated using Terra Mobilis® 2.1 by C. R. Denham and C. R. Scotese; see Wendel & Albert, 1992: 137).

[d] Divergence date used by Wilson et al. (1990), based on the fossil record.

[e] North American Malpighiaceae are here interpreted as representing range expansion from South America.

Nei, 1987). Quasi-ultrametric data may imply selection coefficients very close to neutrality. Remembering that the underlying premise of selective neutrality is the neutral effect of point mutations, nearly clocklike sequence evolution should involve a large proportion of such changes, fixed as effectively neutral substitutions. Such substitutions would be expected to be mainly silent (i.e., synonymous with respect to amino acid[8]), and, with regard to amino acid replacements, functionally conservative (labile). Quasi-ultrametricity in *rbc*L nucleotide sequences is thus an expected manifestation of strong constraints on protein function.[9]

UNIT CHARACTERS AND FUNCTIONAL CONSTRAINTS

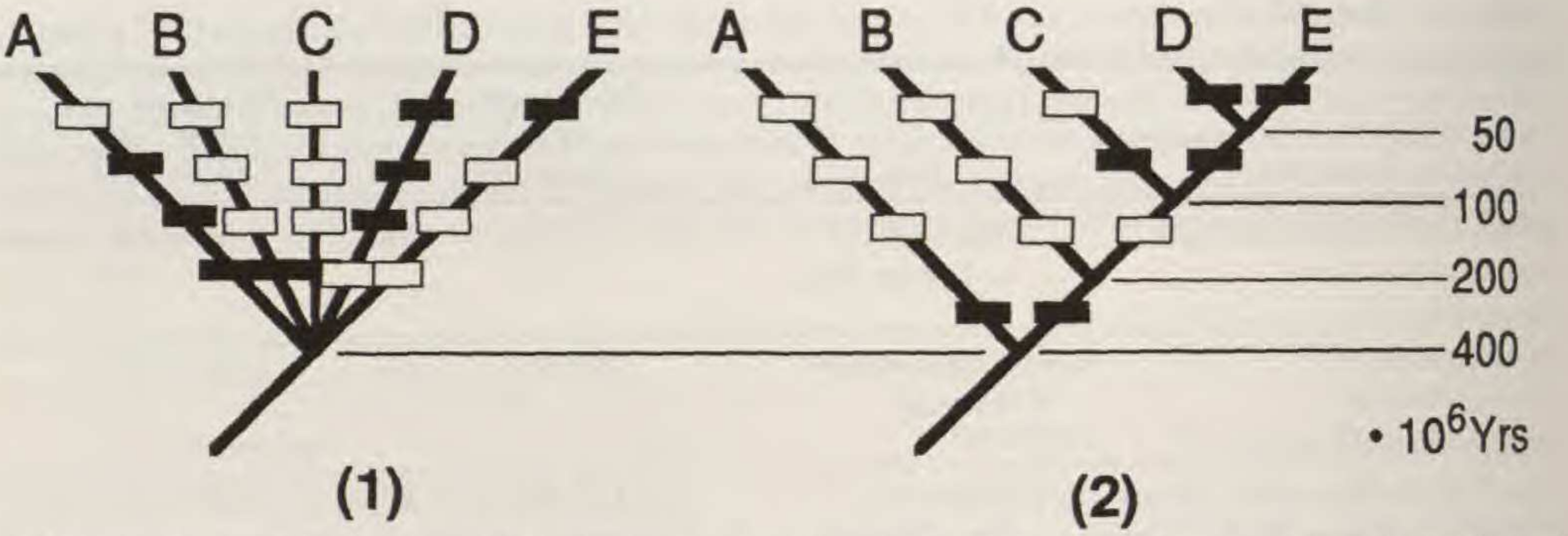
As recently reviewed by Clegg (1993), a number of systematic and evolutionary studies have relied solely on *rbc*L sequence variation. Such analyses make the implicit assumption that *rbc*L nucleotides are independent and potentially informative markers of cladogenetic events. As discussed above with respect to total rates of change, if all branching events under consideration are relatively recent, parsimony analysis may be expected to proceed with a reduced probability of spurious branch attraction because of the absolutely lower expected

[8] See Clegg (1993) on synonymous rates for *rbc*L; note that only *total* substitution rates are relevant to cladistic methods because all informative variation is considered.

[9] Assuming that purifying selection eliminates mutations deleterious to protein function and that f is the fraction of such mutations, the neutral theory may be reformulated as

$$S = (1 - f)\mu$$

where S is the total substitution rate per site and μ is the mutation rate (after Nei, 1987: 52, 411).



FIGURES 1 AND 2. Patterns of historical versus spurious similarity resulting from symmetrically ancient and asymmetrical time-samples. In both cases, time-sample refers to the nodes on these imaginary trees. In (1), all nodes are essentially time-coincident at 400 My, so the "true tree" appears polytomous. In (2), the cladogenetic events indicated occur asymmetrically with respect to time, ranging from 400 to 50 My since divergence. Possible patterns of nucleotide change are indicated by the filled and open rectangles; the former represent unadulterated markers of cladogenetic history, whereas the latter represent spurious character-state similarity resulting, e.g., from multiple nucleotide substitutions. In (1), these patterns of similarity are approximately equal in extent (because of nearly clocklike substitutional behavior) but are in partial conflict with each other; parsimony analysis may include resolutions containing some proportion of ahistorical evidence or even alternatives comprising totally spurious patterns. This might be the expectation if taxa A through E were, e.g., *Isoetes*, *Selaginella*, *Psilotum*, *Equisetum*, and *Angiopteris*. In (2), which approximates the situation in simultaneous studies of sporing and seed plants, the problems of (1) are only partially alleviated. Patterns of convergent similarity between the oldest taxa, A and B, will result in most parsimonious reconstructions that pair these taxa spuriously. As divergence time becomes shallower, the reduced likelihood of multiple changes at sites will insure that D and E are paired historically. Although C is linked with (D, E) by "true" similarity, this relationship may be broken by false similarities between B and C as well as between B, (C, D, E). In summary, comparing only anciently diverged lineages with *rbc*L may suggest patterns of relationship that represent a hopelessly even mixture of historically reliable and nonreliable similarity. Likewise, comparison of ancient and recently diverged clades may have the same problem near the base while being relatively more consistent near the tips. This condition may characterize the *rbc*L-based results shown in this paper.

sequence divergence and relatively lower associated likelihood of character-state parallelism. This "time-sampling" strategy has been employed in circumscribed studies ranging from particular angiosperm groups (e.g., Conti et al., 1993; Kron & Chase, 1993; Rodman et al., 1993) to seed plants as a whole (Chase et al., 1993). Here, a "time sample" refers to the nodes rather than the terminals on an imaginary tree; as such, a time sampling is the collection of absolute and relative timings of underlying cladogenetic events in a data matrix. Of course, the nodes of a cladogram are not discernible a priori to analysis, but their absolute and relative timing may be estimated by external criteria (e.g., the fossil record; cf. Norell & Novacek, 1992).

Initial attempts to analyze time samples beyond angiosperms and other seed plants (i.e., including *rbc*L sequences from sporing plants; Albert et al., 1992b) resulted in cladistic patterns familiar from studies based on ribosomal DNA (rDNA) variation (e.g., monophyletic gymnosperms or combinations of gymnosperm lineages, a seed-plant "root" at the Gnetales, an angiosperm "root" at the monocots; see Troitsky et al., 1991; Zimmer et al., 1989; Hamby & Zimmer, 1992). These results, however, are in conflict with cladistic studies based on morphological characters (see below). Ribosomal RNAs, with their structural and catalytic roles in protein translation, are obviously under enormous functional constraints. Like *rbc*L, rDNAs may also exhibit nearly clocklike substitutional behavior in those positions that are "free" to vary. If the absolute rates of change approximate the low values estimated for *rbc*L, analysis of corresponding time samples might be expected to result in corresponding patterns of homologous and parallel similarity, and therefore similar hierarchial reconstructions (cf. Donoghue & Sanderson, 1992: 347–349).

To gain insight into the topological effects of vastly asymmetrical time samples (see Fig. 2), we have combined *rbc*L information from "bryophytes," "pteridophytes," "gymnosperms," and angiosperms (Table 2). If the substitutional process is effectively clocklike among these taxa, some effects of functional constraints in land-plant *rbc*L evolution should be discernible (as may be spurious branch attractions; see The Rate "Problem," above); we explore this cladistically from both the primary nucleotide data as well as ad hoc nucleotide strings. The *rbc*L data are examined also at the

TABLE 2. *rbc*L sequences used for data transformation and cladistic analysis. These are listed by taxon and by GenBank accession number and/or literature reference where sequence data first appeared. Voucher information, where available, is given by these sources.

Taxon	GenBank accession or literature reference
Conocephalum conicum (L.) Lindb.	Mishler et al., 1994
Lophocolea heterophylla (Schrad.) Dumort.	Mishler et al., 1994
Anthoceros punctatus L.	Mishler et al., 1994
Andreaeobryum macrosporum Steere & B. Murray	Mishler et al., 1994
Ophioglossum engelmannii Prantl	L11058 (J. R. Manhart, in press)
Psilotum nudum (L.) P. Beauv.	L11059 (J. R. Manhart, in press)
Isoetes melanopoda J. Gay & Durieu	L11054 (J. R. Manhart, in press)
Lycopodium digitatum A. Br.	L11055 (J. R. Manhart, in press)
Angiopteris evecta (G. Forst.) Hoffm.	L11052 (J. R. Manhart, in press)
Equisetum arvense L.	L11053 (J. R. Manhart, in press)
Selaginella sp.	L11280 (J. R. Manhart, in press)
Botrychium biternatum (Sav.) Underwood	L13474 (J. R. Manhart, in press)
Taxus × *media*	Chase et al., 1993
Taxodium distichum (L.) Rich.	Soltis et al., 1992
Podocarpus gracilior Pilg.	X58135 (Bousquet et al., 1992)
Ginkgo biloba L.	Chase et al., 1993
Cycas revoluta L.	B. Schutzman, *s.n.*, FLAS, (M. W. Chase, unpublished)
Stangeria eriopus (Kunze) Baill.	Chase et al., 1993
Zamia inermis Vovides, J. D. Reese & M. Vásquez-Torres	L12683 (Chase et al., 1993)
Ephedra tweediana C. A. Mey.	L12677 (Chase et al., 1993)
Welwitschia mirabilis Hook. f.	Chase et al., 1993 (G. R. Furnier)
Gnetum gnemon L.	L12680 (Chase et al., 1993)
Chloranthus japonicus Siebold	L12640 (Chase et al., 1993)
Piper betle L.	L12660 (Chase et al., 1993)
(*Drimys*) *Tasmannia insipida* DC.	L01957 (Albert et al., 1992c)
Calycanthus chinensis Cheng & S. T. Chang	L12635 (Chase et al., 1993)
Eupomatia bennettii F. Muell.	L12644 (Chase et al., 1993)
Magnolia macrophylla L.	Golenberg et al., 1990
Persea americana Mill.	Golenberg et al., 1990
Trochodendron aralioides Siebold & Zucc.	L01958 (Albert et al., 1992c)
Ceratophyllum demersum L.	M77030 (Les et al., 1991) plus nucleotides 1184–1428 from Qiu et al., 1993
Nymphaea odorata Aiton	M77035 (Les et al., 1991) plus nucleotides 1184–1428 from Qiu et al., 1993
Lilium superbum L.	L12682 (Albert et al., 1992a)
Platanus occidentalis L.	L01943 (Albert et al., 1992c)
Caltha palustris L.	L02431 (Albert et al., 1992c)
Dillenia indica L.	L01903 (Albert et al., 1992c)
Chrysolepis (*Castanopsis*) *sempervirens* (Kellogg) Hjelmq.	Chase et al., 1993
Betula nigra L.	L01889 (Albert et al., 1992c)
Casuarina litorea L.	L01893 (Albert et al., 1992c)
Hamamelis mollis Oliv.	L01922 (Albert et al., 1992c)

amino acid level for hierarchic compatibility with the nucleotide and string evidence.

NUCLEOTIDES

The nucleotide is the smallest unit character available in DNA information. With only four states possible at any given site, nucleotide data are subject to parallelism among sequences when the number of changes per site, λ (= rate·time), becomes large. Unlike some morphological characters, nucleotide data are usually analyzed cladistically with no assumed transformation series (i.e., nonadditive steps; Fitch, 1971). For such procedures, Albert et al. (1993) examined the potential for spurious branch attraction under Felsenstein's (1978) simplified four-taxon scenario. State-change probabilities with Jukes-Cantor (Jukes & Cantor, 1969) and Kimura 2-parameter (Kimura, 1980) corrections for multiple changes at sites were considered in addition to observed changes only because of the prospect of reducing character-state parallelisms. All calculations indicated a very small parameter region under which branch attraction could be expected, provided that λ values remained small (i.e., less than approximately 0.1; see Albert et al., 1992a). For quasi-ultrametric data, differences in λ values must principally result from divergence time differences.

The bryophyte lineages examined here could easily be pre-Silurian; the pteridophytes no later than Devonian; the seed-plants appearing by the Carboniferous; the angiosperms by the Cretaceous, followed by their diversification through the Tertiary—a time range potentially spanning 500–5 million years before present. Thus, even without a priori knowledge of precise divergence times, it is reasonable to approximate upper and lower λ-bounds from this range and our estimates of total sequence divergence. The mean rate for woody taxa (Table 1), averaged for single lineages by halving the divergence value, is approximately 1.0×10^{-10} nucleotide substitutions per site per year. Similarly, the estimates for herbaceous taxa (Wendel & Albert, 1992) range between $2.5–3.5 \times 10^{-10}$. Assuming that bryophytes and pteridophytes fall into the range $1.0–3.5 \times 10^{-10}$ as well, λ values are estimated to lie between 0.05–0.175 (500 My) and 0.0005–0.00175 (5 My). On a four-taxon tree, some combinations of these values would yield spurious branch attractions (see Albert et al., 1993). Here, we are working with 40 taxa and a greater potential for inconsistent results (see Penny et al., 1991).

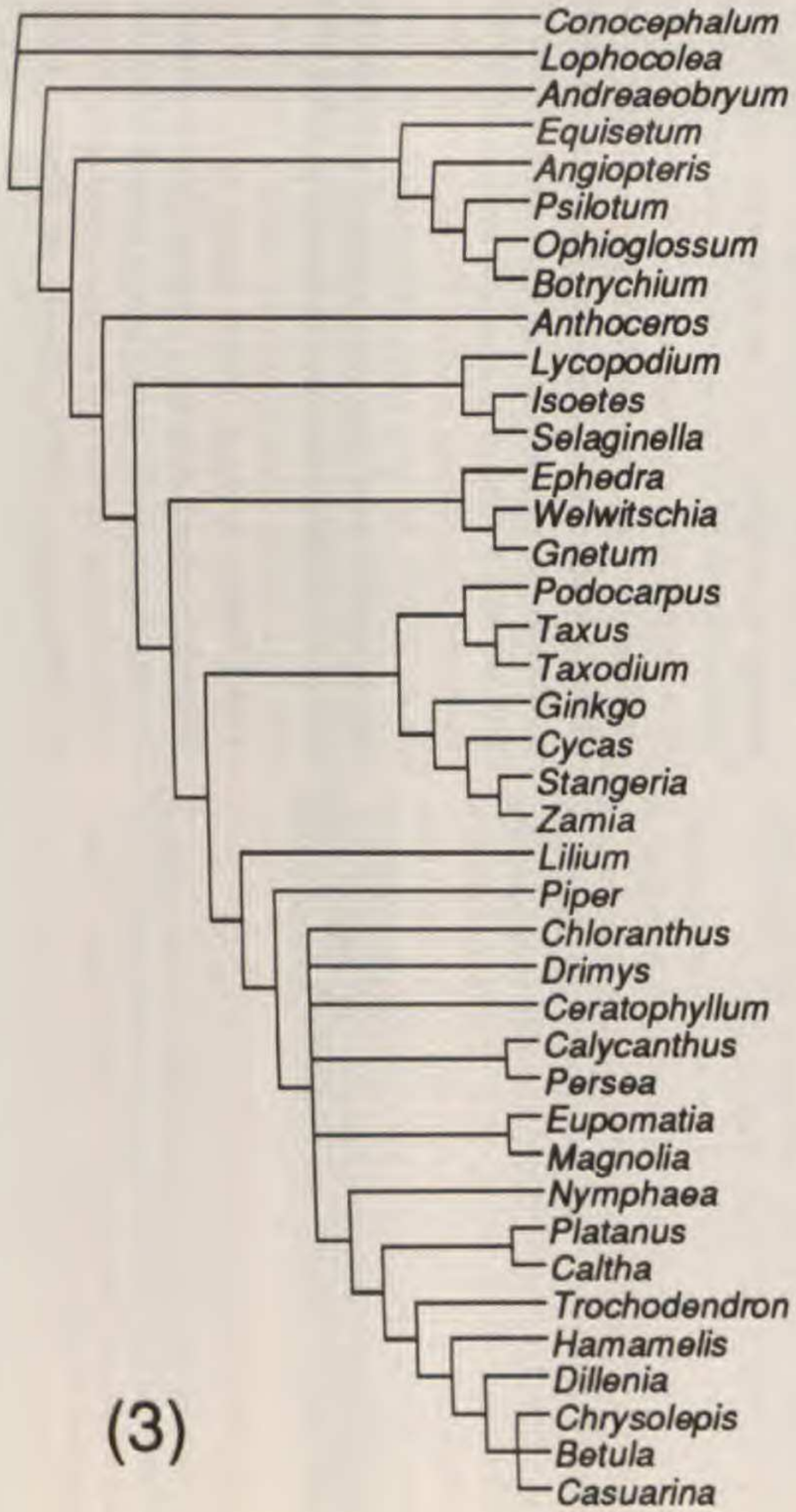
Data analysis. Nucleotide sequences (unambiguously aligned by sight and excluding the 30 5′-most positions, which incorporated only primer information for some taxa; Table 2) were analyzed with PAUP 3.1.1 (Swofford, 1993) using the Fitch criterion (Fitch, 1971; cf. Albert et al., 1993) with ACCTRAN (accelerated transformation) optimization (Farris, 1970; Swofford & Maddison, 1987). The heuristic search option was used with 100 random replicates of data addition sequence, COLLAPSE, MULPARS, and TBR (tree bisection-reconnection) branch-swapping. The consistency and retention indices (*C* and *R*, respectively; Kluge & Farris, 1969; Farris, 1989a) were also calculated. Five hundred fifteen nucleotide positions showed patterns of similarity among taxa.

Eight equally parsimonious cladograms were found ($C = 0.362$ (including all data), $R = 0.523$). The strict and combinable component consensus trees (Bremer, 1990) were identical (see Fig. 3). All trees indicate that (i) hornworts are nested inside the tracheophyte clade, (ii) lycopods rather than ferns plus *Equisetum* represent the sister group to seed plants, (iii) Gnetales represent the sister group of all other seed plants, (iv) conifers, *Ginkgo*, and cycads form the monophyletic sister group to angiosperms, and (v) monocots are basalmost in the angiosperms, followed by *Piper*. Characteristics (iii) and (iv) are shared with the rDNA analysis of Hamby & Zimmer (1992) but not with the morphological analyses of Crane (1985), Doyle & Donoghue (1986, 1992), Loconte & Stevenson (1990), and Nixon et al. (1994). Characteristic (i) is in conflict with both morphological and molecular cladistic studies (Mishler & Churchill, 1985; Mishler et al., 1994, this issue). Characteristic (ii) contrasts both with morphological data (Bremer, 1985) and with the chloroplast genome structural findings of Raubeson & Jansen (1992) that link all tracheophytes except the lycopods, which have the plesiomorphic (i.e., liverwortlike) state. Characteristic (v) contrasts with the results of morphological (Donoghue & Doyle, 1989; Loconte & Stevenson, 1991; Taylor & Hickey, 1992) and some rDNA (Hamby & Zimmer, 1992; cf. Zimmer et al., 1989) analyses.

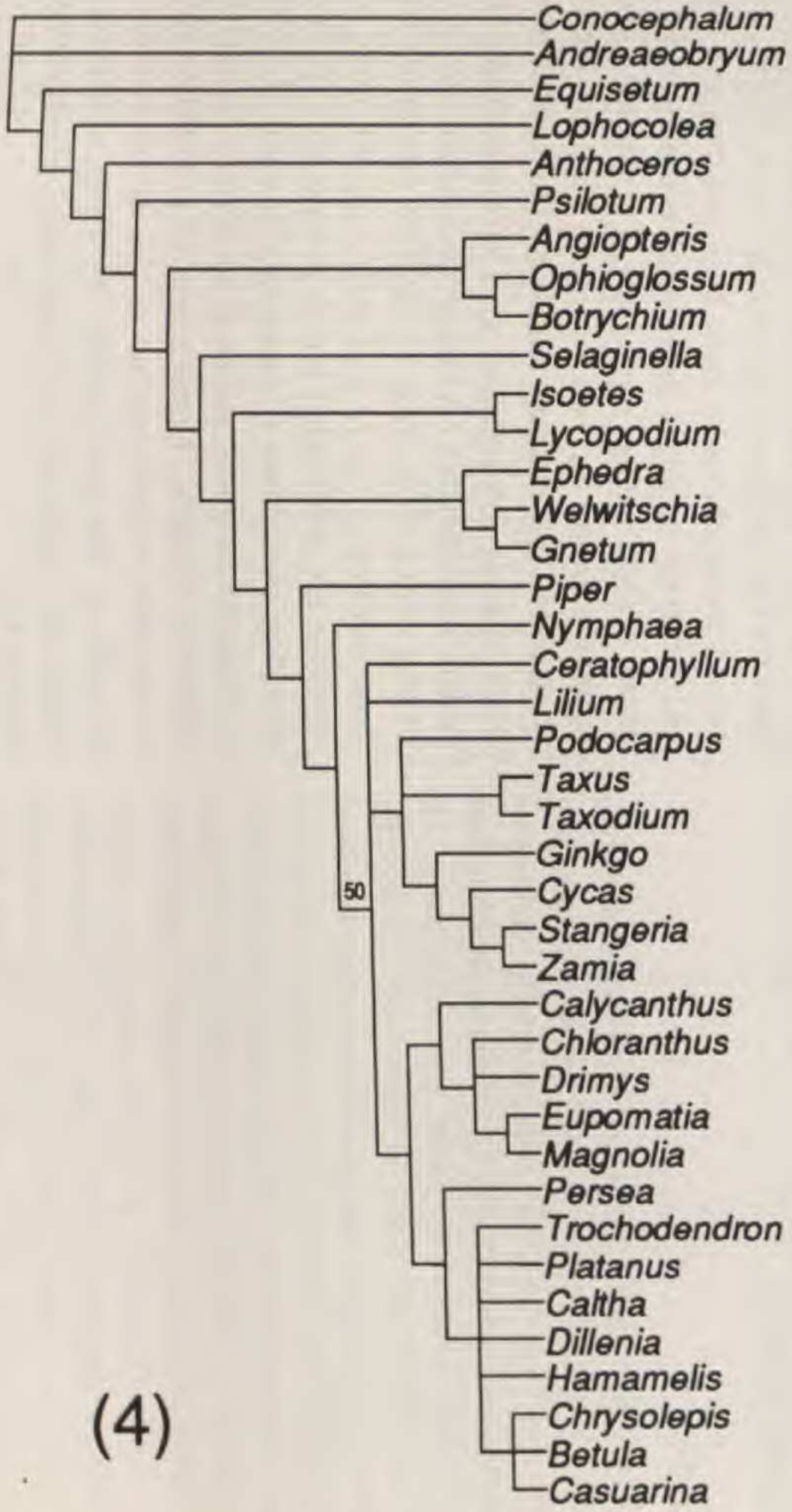
Function and phylogeny. Needless to say, not all of the above observations can represent the truth about land-plant history. The groups found in the nucleotide-based parsimony analysis (Fig. 3) may well reflect historical reality, but the nature of that reality could be other than strictly phylogenetic. From our argument about nearly clocklike rates and the functional constraints that may produce them, it is reasonable to suppose that some or even all of the branchings depicted in Figure 3 may reflect primarily spurious similarities rather than phylogenetic homologies. We have assessed possible constraints on *rbc*L evolution by examining the amino acid changes implied on the internal

→

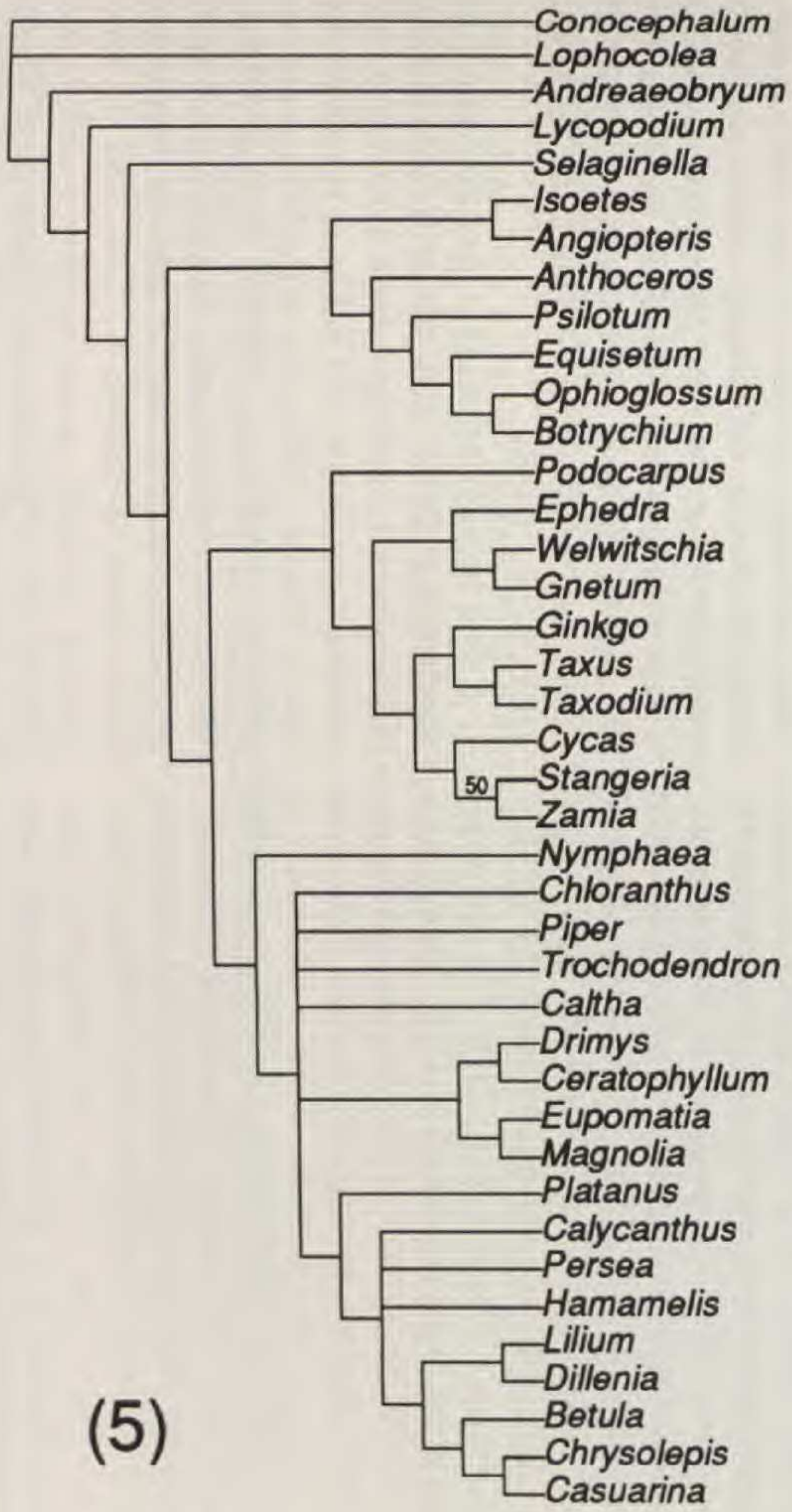
FIGURES 3–5. Combinable component consensus trees summarizing the results of parsimony analyses of *rbc*L evidence as (3) nucleotide, (4) string, and (5) amino acid data. For (3), the strict consensus is identical; for (4) and (5), the single combinable components are indicated by the percentage of most parsimonious trees that resolve what would otherwise be polytomies. Implications of the different topologies are discussed in the text.


Conocephalum
Lophocolea
Andreaeobryum
Equisetum
Angiopteris
Psilotum
Ophioglossum
Botrychium
Anthoceros
Lycopodium
Isoetes
Selaginella
Ephedra
Welwitschia
Gnetum
Podocarpus
Taxus
Taxodium
Ginkgo
Cycas
Stangeria
Zamia
Lilium
Piper
Chloranthus
Drimys
Ceratophyllum
Calycanthus
Persea
Eupomatia
Magnolia
Nymphaea
Platanus
Caltha
Trochodendron
Hamamelis
Dillenia
Chrysolepis
Betula
Casuarina

(3)


Conocephalum
Andreaeobryum
Equisetum
Lophocolea
Anthoceros
Psilotum
Angiopteris
Ophioglossum
Botrychium
Selaginella
Isoetes
Lycopodium
Ephedra
Welwitschia
Gnetum
Piper
Nymphaea
Ceratophyllum
Lilium
Podocarpus
Taxus
Taxodium
Ginkgo
Cycas
Stangeria
Zamia
50
Calycanthus
Chloranthus
Drimys
Eupomatia
Magnolia
Persea
Trochodendron
Platanus
Caltha
Dillenia
Hamamelis
Chrysolepis
Betula
Casuarina

(4)


Conocephalum
Lophocolea
Andreaeobryum
Lycopodium
Selaginella
Isoetes
Angiopteris
Anthoceros
Psilotum
Equisetum
Ophioglossum
Botrychium
Podocarpus
Ephedra
Welwitschia
Gnetum
Ginkgo
Taxus
Taxodium
Cycas
50
Stangeria
Zamia
Nymphaea
Chloranthus
Piper
Trochodendron
Caltha
Drimys
Ceratophyllum
Eupomatia
Magnolia
Platanus
Calycanthus
Persea
Hamamelis
Lilium
Dillenia
Betula
Chrysolepis
Casuarina

(5)

branches of one of the eight equally most-parsimonious trees (Appendix I). As summarized in Table 3, over 84% of the inferred nucleotide substitutions on internal branches are silent with regard to amino acid identity. The percentage of nucleotide changes incurring functionally labile amino acid replacements (judged using the PAM-250 log-odds matrix of Dayhoff et al., 1978: 352; see Table 3) amount to an additional ≈ 13%. Viewed as a whole, 97.5% percent of all synapomorphous nucleotide changes are expected to have little or no effect on protein function. With a maximum of only 2.5% of these changes incurring non-labile amino acid replacements of potential structural/functional distinction (see Table 3), *rbc*L sequences appear heavily burdened by forces leading to functional conservation.[10] Thus, the challenge for land-plant cladistics is to determine how strongly functionally constrained variation may also reflect phylogenetic patterns.

STRINGS

The ideal "unit" character in phylogenetic analysis is one that truly evolves as an independent unit, meaning one that independently undergoes transformations from one condition to another that are hierarchically correlated (i.e., congruent; cf. Farris, 1969) with those of other such characters. For molecular data, this may often be the individual nucleotide, but possibly also a contiguous length of DNA in an insertion/deletion event, several non-contiguous nucleotide positions that are functionally associated (e.g., because of higher order RNA or protein structure), a unique codon for a functionally constrained amino acid, or a whole chromosome in a karyological change. It is of course difficult to assess such possibilities a priori, but it is nonetheless important to begin to develop methods to examine the issue empirically.

We have thus examined some means by which the functional/phylogenetic evidence manifest in a given set of *rbc*L sequences might be represented by data forms other than nucleotide positions and their character states. The nucleotide is indeed the smallest unit character in *rbc*L evidence, but it is not necessarily the most informative nor most consistent. First, nonadditive optimization of multistate characters may restrict potential topological resolution (e.g., a 4-state, nonadditive character can have minimum homoplasy if optimized as three autapomorphies). Additionally, direct analysis of nucleotide sequences from protein-coding genes ignores constraints imposed both by the genetic code and protein function; codon positions may be both intra- and inter-correlated (Fitch & Markowitz, 1970; Fitch, 1986).

A data transformation that may overcome these shortcomings stems from the early comparison of oligonucleotide catalogues (and even whole chromosomes; see Farris, 1978; Fox et al., 1980; Bremer & Bremer, 1989) prior to the DNA sequencing revolution: production of ad hoc nucleotide strings. Our procedure (analogous to generating mapped restriction site data) may be outlined thus: (i) generate strings of random A, T, G, and C content varying randomly in size between 6 and 21 base pairs (so that a minimum and maximum of two and seven codons are included), (ii) scan *rbc*L sequence data for the presence/absence of given strings, (iii) record recognitions by both base position and taxon, (iv) treat multiple positional recognitions by a given search string separately, (v) treat all recognitions found in two or more taxa as binary characters for cladistic analysis (sequences that have missing information at a string position are coded accordingly). Another procedure for producing string data from nucleotide sequences has been developed by J. S. Farris (unpublished); sequences are subdivided into a prespecified number of string characters ("supersites"), each of which is assigned as many states as necessary to explain observed variation. Farris's method guarantees both a complete transformation of the entire sequence as well as the non-overlap of string characters, unlike the approach used here (see below and Appendix II).

The net effect of transforming sequences into strings is twofold: (i) it incorporates more information (in terms of nucleotides or codons spanned) in a larger unit character, and (ii) decreases the probability that independent gains of the same character-state are represented in data matrices (although, in parsimony analyses, binary characters are more subject to spurious branch attraction than are nonadditive multistate characters; Albert et al., 1993). As with mapped restriction site data, the probabilities of gain versus loss of a recognition string are highly asymmetrical, with parallel gains the least likely transformation series (Templeton, 1983; DeBry & Slade, 1985; Albert et al., 1992a). Therefore, string data may contain historical markers much less likely to engage in branch attraction (which occurs because of accumulated parallelisms; cf. Felsenstein, 1978; Hendy & Penny, 1989;

[10] Patterns of codon usage intrinsic to the primary nucleotide matrix are also suggestive of functional constraints; these are discussed in a separate paper (Albert, Backlund & Bremer, in press).

TABLE 3. Analysis of character support for internal branches of tree #1 (of 8) from the nucleotide analysis. "Node" refers to the node numbers on the reference tree of Appendix I. "# changes" refers to the total number of nucleotide changes optimized onto a branch. "Constant" indicates that the nucleotide site belongs in a codon position that codes for the same amino acid throughout the entire matrix. "No change" indicates that the nucleotide site belongs in a codon position that codes for two or more amino acids throughout the matrix, but that the particular change indicated at this node does not cause a change in amino acid sequence. "Labile" means that the inferred change in amino acid due to the observed change in nucleotide sequence is likely to happen by random chance or better (according to the PAM-250 log-odds matrix of Dayhoff et al., 1978: 352). "Potentially nonlabile" indicates that at least one of the potential amino acid changes inferred from a particular nucleotide position is not likely to happen by random, but that there also are some changes in the same character that are likely to happen by random chance or better. "Nonlabile" means that all inferred acid changes (often only one) occur at less than random chance.

Node	# changes	Constant	No change	Labile	Potentially nonlabile	Nonlabile
78-77	42	22	4	8	5	3
77-76	24	13	6	4	0	1
76-71	27	13	9	3	2	0
71-70	29	19	9	1	0	0
70-42	40	24	11	5	0	0
42-41	33	26	5	1	0	1
70-69	42	17	16	8	0	1
69-66	29	21	8	0	0	0
66-48	34	15	13	5	0	1
48-44	25	10	12	2	0	1
44-43	29	19	8	2	0	0
48-47	15	7	8	0	0	0
47-46	24	14	7	3	0	0
46-45	11	4	4	3	0	0
66-65	56	34	15	7	0	0
65-64	26	13	10	3	0	0
64-63	18	11	6	1	0	0
63-54	5	2	0	3	0	0
54-53	4	3	0	1	0	0
53-51	10	3	1	5	1	0
51-49	9	4	2	3	0	0
51-50	8	2	1	5	0	0
53-52	11	5	2	4	0	0
63-62	16	11	5	0	0	0
62-61	14	6	7	1	0	0
61-59	8	2	4	2	0	0
59-58	17	8	5	4	0	0
58-57	13	6	4	3	0	0
57-56	33	20	6	7	0	0
56-55	6	3	2	1	0	0
61-60	8	5	2	1	0	0
69-68	58	29	18	8	3	0
68-67	45	24	17	4	0	0
76-75	34	20	7	4	0	3
75-74	38	23	12	2	1	0
74-73	45	28	14	3	0	0
73-72	65	43	12	9	1	0
Σ	951	529	272	126	13	11
	100.00%	55.63%	28.60%	13.25%	1.37%	1.16%
		84.23%				
		97.48%				

Albert et al., 1992a, 1993) and much more likely to contain "blocks" of evolutionarily correlated information. Nevertheless, this information could be functionally constrained, as with primary nucleotide data. This possibility can be studied similarly by examining inferred amino acid changes on cladograms; each string character is easily traced to its recognized codons and component nucleotides.

Data analysis. One thousand random strings were generated for evaluation (see Appendix II). After scanning the 40 *rbc*L sequences, 193 positionally distinct string recognitions were recorded (mostly from small strings, the largest being from a 15-mer; see Appendix II). Of these, 112 identified two or more taxa. As there was no control in our procedure for string overlap, a number of string recognitions are non-independent with respect to nucleotides identified (see Appendix II). Therefore, our string data carry an experimental bias similar to what could occur with restriction site data representing mapped cleavage points for several endonucleases. The "supersites" string transform (J. S. Farris, unpublished) avoids this difficulty entirely, and if modified for the production of presence/absence data, would be identical to our intent but superior in execution. Nevertheless, our string data should suffice to explore biological non-independence of nucleotides (functional constraints); in fact, partial replication of nucleotide "blocks" could enhance detection of conserved regions. Cladistic analysis of the string characters was performed under the Wagner criterion (Kluge & Farris, 1969; Farris, 1970; see Albert et al., 1992a) using the same program and options mentioned previously; 165 equally parsimonious trees were found ($C = 0.381$ (including all data), $R = 0.524$). The combinable component consensus tree differs from the strict by only one component (see Fig. 4).

The string data provide a different resolution of land-plant relationships than the nucleotide sequences (Figs. 3, 4). Notable differences include (i) *Equisetum* placed among the bryophytes, (ii) paraphyly of *Psilotum* + ferns and paraphyly of lycopods, (iii) sister-group status of Gnetales to angiosperms (with *Piper* basalmost), and (iv) paraphyly of angiosperms to conifers + (*Ginkgo*, cycads). Characteristics (i) and (iv) are in total conflict with other results (listed under Nucleotides, above), whereas (ii–iii) are not.

Function and phylogeny. It could be argued that cladograms produced from string-transformed data are better phylogenetic representations than those derived from nucleotides because the unit character is substantially less subject to parallel gains (see above). However, this attribute is distinct from the nature of the history conserved by string data; whole functional units may be incorporated into single characters. Gross differences in tree topology (including paraphyly of angiosperms) may simply result from different representations of functional and phylogenetic history in string versus nucleotide data forms.

We have studied possible functional constraints on *rbc*L evolution (as above) by examining the inferred amino acid changes on the internal branches of one of the 165 equally most-parsimonious string trees (Appendix II). Striking differences from the nucleotide-based analysis (Table 3) are shown in Table 4: only 45% of string transformations (changes in underlying nucleotide sequence) are silent with regard to amino acid identity (versus ca. 84% in the nucleotide analysis, a decrease by half), and functionally labile amino acid replacements amount to an additional 25% (versus ca. 13% in the nucleotide analysis, a relative increase). Thus, 70% of underlying nucleotide changes appear to be functionally neutral, whereas non-labile amino acid replacements amount to a maximum of 28% (an additional 2.1% are ascribed to internal stop codons, which may result from sequencing errors). This greater number of presumably functional changes in underlying nucleotides does indicate a greater chance that functional associations among particular nucleotides may bias tree construction.

The different substitutional patterns between nucleotide and string data can be explained by inherent properties of the latter. Each string recognition shared by two or more sequences comprises much more inclusive and conservative information than shared nucleotide identity at a given site. From our previous arguments about functional constraints in *rbc*L sequence evolution (see The Rate "Problem" and Nucleotides, above), the majority of string recognitions are expected to identify functionally conserved nucleotide motifs. The proportional reduction in discernible silent substitutions on the nucleotide level is likely due to the increased size of the functional units compared; with a 6 base-pair string, the chance of observing a non-silent change is at least six times greater than for a single nucleotide position. The proportional increase in labile amino acid replacements can be explained through similar reasoning; if a string recognition identifies a functionally conserved motif, the larger the motif, the greater the likelihood that functional preservation need not require exact

TABLE 4. Analysis of character support for internal branches of tree #100 (of 165) from the string analysis. "Node" refers to the node numbers on the reference tree of Appendix II. "# changes" refers to the total number of string changes optimized onto a branch. "Constant" indicates that the string identifies codon positions that code for the same amino acid throughout the entire matrix. "Labile" means that the inferred change in amino acid due to the observed change in string recognition is likely to happen by random chance or better (according to the PAM-250 log-odds matrix of Dayhoff et al., 1978: 352). "Potentially nonlabile" indicates that at least one of the potential amino acid changes inferred from a particular string recognition is not likely to happen by random, but that there also are some changes in the same character that are likely to happen by random chance or better. "Nonlabile" means that all inferred amino acid changes (often only one) occur at less than random chance. "Internal stop" refers to string recognitions that identify internal stop codons, which may be sequencing artifacts.

Node	# changes	Constant	Labile	Potentially nonlabile	Nonlabile	Internal stop
77-76	7	3	3	1	0	0
76-75	7	5	2	0	0	0
75-74	6	1	1	2	1	1
74-73	4	4	0	0	0	0
73-72	4	4	0	0	0	0
72-42	4	2	2	0	0	0
42-41	9	4	3	1	1	0
72-71	4	1	1	2	0	0
71-70	6	4	2	0	0	0
70-43	5	1	1	3	0	0
70-69	2	2	0	0	0	0
69-66	7	3	1	3	0	0
66-65	4	1	1	0	2	0
65-51	3	1	1	0	0	1
51-50	3	1	1	1	0	0
50-49	6	4	2	0	0	0
49-48	8	3	3	0	1	1
48-44	6	3	1	0	2	0
48-47	4	1	0	2	1	0
47-46	8	2	3	1	2	0
46-45	3	2	0	0	1	0
65-64	4	1	2	0	1	0
64-55	3	1	1	0	1	0
55-54	2	0	1	0	1	0
54-52	1	1	0	0	0	0
54-53	3	2	0	0	1	0
64-63	1	1	0	0	0	0
63-62	4	1	1	1	1	0
62-61	2	0	0	0	2	0
61-56	2	0	1	1	0	0
60-57	2	1	0	0	1	0
60-59	5	1	2	1	1	0
59-58	2	1	1	0	0	0
69-68	4	1	1	1	0	1
68-67	10	6	1	1	2	0
Σ	155	69	39	21	22	4
	100.00%	44.51%	25.16%	13.55%	14.19%	2.08%
		69.67%		27.74%		

amino acid identity. Strings recognizing regions of non-labile change, indicating potentially radical changes in structure and function among taxa, may represent another class of conserved information. Again, these are probably found in greater proportion because of the larger size of the unit characters. Rather than being conserved because of functional constraints (as above), such recognitions may identify conserved markers for historical groups. Such changes may or may not have drastic physiological effects (see Hudson et al., 1990, on *rbc*L ; cf. Perutz & Lehman, 1968; Nei, 1987: 270–271), but they could be of similar phylogenetic utility as chloroplast DNA rearrangements (e.g., Jansen & Palmer, 1987; Palmer et al., 1988; Bruneau et al., 1990; Lavin et al., 1990; Downie & Palmer, 1992; Downie et al., 1991; Raubeson & Jansen, 1992) if well characterized in relation to the crystal structure of the large-subunit protein (Chapman et al., 1988; Andersson et al., 1989; cf. Clegg, 1993).

AMINO ACIDS

Because *rbc*L nucleotide substitutions approximate a clock hypothesis (see The Rate "Problem," above), amino acid changes are expected to conform to the neutral hypothesis of molecular evolution (see Nei, 1987: 47–59, 409–412), although we do not directly address this issue here. Direct inference of trees can proceed from amino acids (yet another transformation of the same primary evidence). One limitation of using the amino acid sequences themselves is the "factoring-out" of all synonymous variation at the nucleotide level; this again may make it more likely that functional associations among characters may bias tree construction. Topological resolution may also be limited because amino acid data is optimized nonadditively (Fitch, 1971) and more than four states could be available for given characters (in the *rbc*L sequences examined here, the maximum is six states at four different positions). Nevertheless, the greater the number of character states, the lower the probability of character-state parallelism and spurious branch attraction (Albert et al., 1993). It could thus be argued that amino acid data might be more suitable for bridging large evolutionary time gaps, given a roughly constant rate of substitution combined with ignorance of potentially multiple synonymous nucleotide changes. Hence, we evaluated the amino acid data for hierarchic compatibility with the results of the nucleotide and string analyses.

Data analysis. After "translating" the 40 *rbc*L sequences, 66 (out of the 476) amino acid positions identified two or more taxa. Cladistic analysis of these characters was performed under the Fitch criterion (Fitch, 1971) using the same program and options mentioned previously; 104 equally parsimonious trees were found ($C = 0.567$ (including all data), $R = 0.554$). The combinable component consensus tree preserved one more component than the strict (see Fig. 5).

The amino acid data provide yet another resolution of land-plant relationships (cf. Figs. 3, 4): (i) lycopods are polyphyletic, with *Isoetes* sister to *Angiopteris*, (ii) *Anthoceros* is embedded among fern allies, (iii) gymnosperms as a whole (with conifers polyphyletic) are the monophyletic sister group to angiosperms (with *Nymphaea* basalmost), and (iv) *Lilium* is sister to *Dillenia*. Except for gymnosperm monophyly as hypothesized from rDNA data (see Troitsky et al., 1991) these characteristics are in total conflict with all previous studies (listed under Nucleotides, above).

From the arbiter of congruence, large-subunit amino acid data are no more appropriate for bridging gaps in asymmetric time samples than nucleotide or string data. As argued above, the clocklike behavior of *rbc*L nucleotide substitution is expected to obtain also in the translated amino acid data; thus, λ values for amino acid changes (and so the likelihood of spurious branch attraction) should also be sensitive to differences in divergence times.

Function and phylogeny. Amino acid changes in *rbc*L are apparently subject to strong functional constraints (see Nucleotides and Strings, above). One could argue that amino acid data is less subject to the "noise" of neutrality, i.e., multiple silent changes at given nucleotide positions. However, selective neutrality may be roughly maintained by labile amino acid replacements, which could similarly "wobble" back and forth across evolutionary time. Only a small percentage of individual amino acids appears to be involved in function-changing evolutionary events (see Nucleotides, above).

PENULTIMATE CONCLUSIONS

We have demonstrated the problematic, functionally constrained nature of *rbc*L markers currently being used for phylogeny estimation by many workers. Three transformations of the same evidence produced discordant cladistic topologies and substantial incongruence with previous morphological cladistic results. Of course, we do not suggest that the growing *rbc*L database be abandoned. Rather, we suggest (as will be elaborated below) that all investigators involved with *rbc*L or other gene data take heed of standard and powerful cladistic procedures for discriminating cladistic history (homology) from homoplasy (functional parallelism and reversal).

Total Evidence and Character Congruence

(i) on characters

Every character in a data matrix showing similarity between two or more taxa is optimized under parsimony as a discrete and independent piece of information. This holds whether or not the character represents a single taxic homology or only a portion of one (which is the case with correlated or contingent characters). A taxic homology used in parsimony analysis is expected to have a single functional history (even if this history changes over time; see Riedl, 1978; Donoghue, 1989; Donoghue & Sanderson, 1992); its cladistic utility (i.e., optimization as synapomorphy or homoplasy) is tested at maximum parsimony along with all other characters in a matrix. From our argument about shared functional history (constraints) in the evolution of *rbc*L, one might be tempted to equate a given taxic homology (e.g., nuclear versus cellular endosperm development) with the entire *rbc*L gene. However, unlike a given taxic homology, *rbc*L is composed of multiple, discrete points of information, that is, its ca. 1428 nucleotides. To a parsimony algorithm, each of these data points is equivalent to the single, nonadditive taxic homology statement "functional pollen unit in the Orchidaceae: monad, tetrad, massula, or pollinium," whatever its underlying complexity.

Hence, some workers have found cladistic philosophy and methodology at an impasse. For example, it has been argued that gene information could be combined with other characters either through multistate recoding of gene trees (Doyle, 1992) or through analysis of component compatibility among separately produced cladograms (Page, 1993). Legitimate concern over potentially separate phylogenetic histories led to these suggestions, but we argue below that both approaches unnecessarily restrict the information content of cladistic hierarchies, a feature fundamental to the superiority of parsimony methods (see Farris, 1979, 1983); in fact, parsimony itself arbitrates the supposed analytical quandary.

(ii) on evidence

For cladistic analysis, evidence is the body of available information that shows patterns of similarity among terminals. A specific set of evidence may be expressed in different forms; we have shown this property through different data transformations of the *rbc*L gene (above). Approaches that combine evidence in the form of tree components do so at the cost of information content (for recent debate on this issue, see Jones et al., 1993; Nelson, 1993; Barrett et al., 1993; De Queiroz, 1993). In fact, acceptance of parsimony as the arbiter of synapomorphy and homoplasy seems methodologically counterintuitive to component combination, which does not directly use such information (see Doyle, 1992; Page, 1993). Parsimony, acting over all evidence, will provide estimates of congruence among character-state patterns while *minimizing* ad hoc assumptions (Farris, 1983). For example, some characters from a multigene family (gene duplication being part of the functional burden) may not show congruence with the body of retained synapomorphy because of paralogous histories (Fitch, 1970). Nevertheless, analysis of "total" evidence (sensu Kluge, 1989) gives each data point the opportunity both to affect hierarchy directly and to be diagnosed objectively, which is not the case when evidence is decomposed a priori and later combined or reconciled (cf. Doyle, 1992; Page, 1993). In conclusion, although a functionally constrained DNA sequence like the *rbc*L gene may appear to deserve the same rank as a given morphological character, it is more evidence-rich, and all of this evidence can be examined for hierarchic correlation (sensu Farris, 1969) with other data.

(iii) an example

The extent to which *rbc*L evidence shows hierarchic correlation with other evidence should provide an objective measure of its freedom from biasing functional considerations, and consequentially, its phylogenetic utility. In this context, we examined character interaction between *rbc*L evidence and the primarily morphological seed-plant matrix of Nixon et al. (1994). Using the set of functional histories in the morphological matrix as a "constant," we tested the ability of different *rbc*L data forms (i.e., nucleotides, strings, and amino acids) to produce a unified representation of the same evidence. Two different sets of experiments were performed: (i) analyses including fossil taxa for which *rbc*L evidence is lacking (and therefore coded as missing data), and (ii) analyses of data for extant taxa only (the intersection of available evidence). To measure character congruence, we have used the retention index: the proportion of congruent similarity (i.e., synapomorphy) in a data matrix that is retained at maximum parsimony (see Farris, 1989a, b, 1991). Although retention is not directly comparable among different data matrices (see Goloboff, 1991), each matrix within our respective sets of experiments shares the same "constant." Additionally, each data transform of *rbc*L

TABLE 5. Homoplasy and character congruence statistics for total evidence analyses comprising morphological (Nixon et al., 1994; matrix version as of 8 November 1993) and *rbc*L data. Consistency (over all data) and retention indices are listed (see text), along with the number of trees found (see Figs. 6–8). For comparisons involving both fossil and extant taxa, 101 morphological similarities are relevant (symbolized by "N"); for extants only, there are 96 (symbolized by "N_{ex}"). The numbers of relevant similarities for each *rbc*L data transform (nucleotides, strings, amino acids) are given in the text. For analyses including fossil taxa, *rbc*L evidence was represented as missing (i.e., "?").

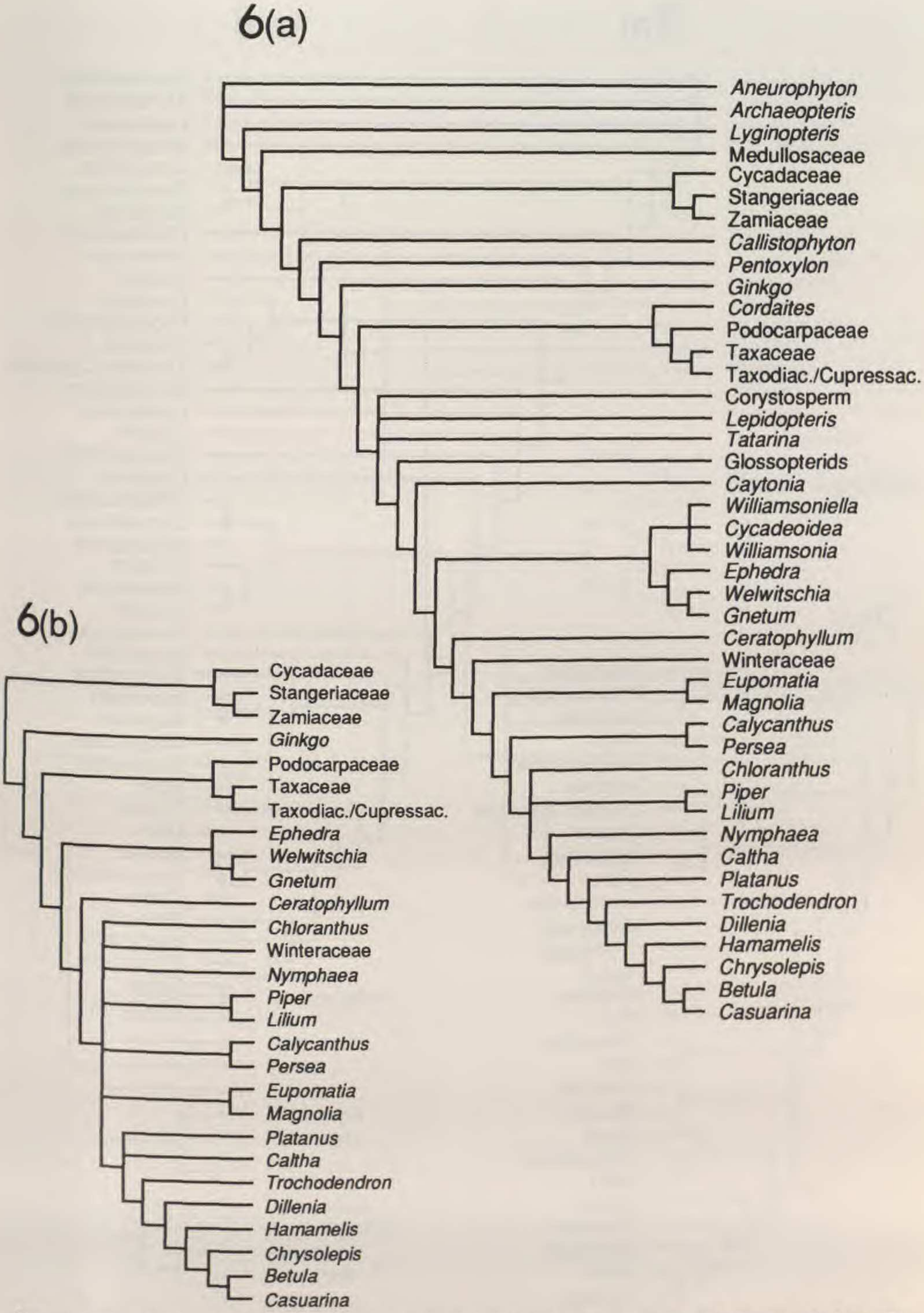
	Consistency	Retention	# Trees
Fossil plus extant taxa			
N + nucleotides	0.450	0.625	44
N + strings	0.402	0.685	22
N + amino acids	0.467	0.710	309
Extant taxa only			
N_{ex} + nucleotides	0.464	0.601	3
N_{ex} + strings	0.442	0.641	7
N_{ex} + amino acids	0.518	0.670	24

is assumed to be evidentially equivalent until shown otherwise (this assumption is obviously weaker for the string data, as they do not represent a completely saturated transformation of the nucleotide sequences). Finally, we do not use retention to suggest which analysis(es) may be "better."

The characters and cladistic reconstructions for living and fossil seed plants are described elsewhere (Nixon et al., 1994). We used the same parsimony methods outlined above to examine six combined matrices comparing all versus extant-only taxa and nucleotide/string/amino-acid *rbc*L data in all combinations. Consistency and retention indices for each analysis are reported in Table 5, and topological results are summarized in Figures 6–8. Character congruence, as measured through retention, is similar in magnitude (range < 0.1) across each set of experiments. Although topological resolution and component placements differ somewhat with respect to the *rbc*L data form used (Figs. 6–8; see Nixon et al., 1994), the *rbc*L evidence appears to be making a consistent statement along with the morphological evidence.

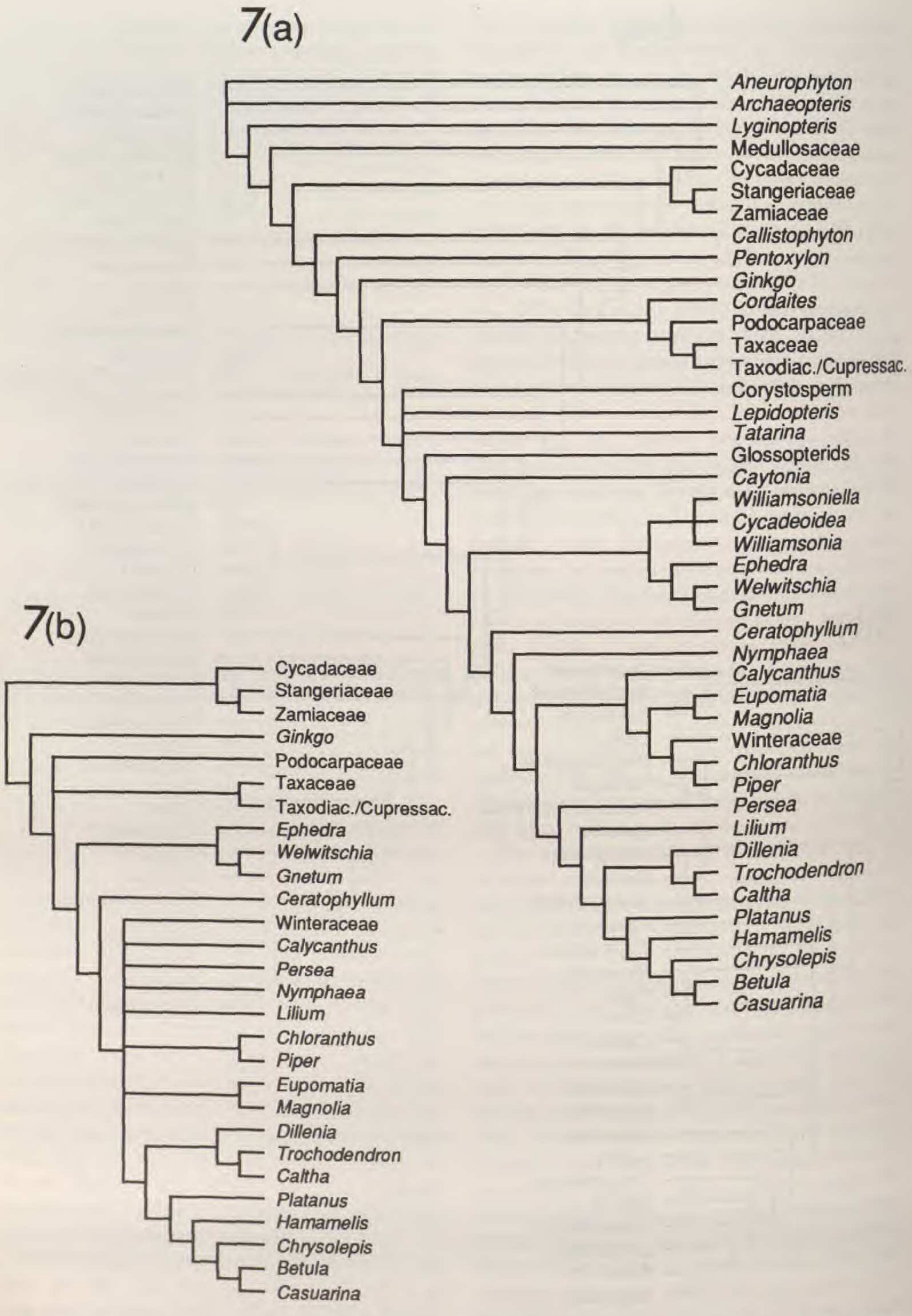
With respect to extant taxa, monophyletic cycads are the most topologically ancestral in all analyses including fossils (Figs. 6a–8a). *Ginkgo* appears either external to *Cordaites* plus conifers (Figs. 6a, 7a) or monophyletic with these taxa (Fig. 8a). In extant-only analyses, *Ginkgo* similarly intercalates between cycads and conifers (Figs. 6b, 7b) or remains sister to conifers (Fig. 8b). Conifers themselves are monophyletic in most combined analyses (Figs. 6a, b, 7a, 8a, b), but are partially unresolved in the extant-only analysis with string data (Fig. 7b). Every analysis resolves the Gnetales and Bennettitales as sister to the angiosperms. *Ephedra* is uniformly sister to *Gnetum* plus *Welwitschia*, but resolution within Bennettitales is provided only in the combined analysis with amino acid data (Fig. 8a). *Ceratophyllum* is placed sister to all other angiosperms (see Les, 1988; Chase et al., 1993; Qiu et al., 1993) in the combined nucleotide and string analyses (Figs. 6a, b, 7a, b), but not in the combined amino acid analyses (Fig. 8a, b), where it either nests well within angiosperms (sister to *Chloranthus*; Fig. 8a) or remains unresolved (Fig. 8b). Indeed, relationships within the angiosperms are the least stable among the combined data analyses. Woody magnoliids occupy the basalmost branches in Figure 6a, whereas the "paleoherb" taxon *Nymphaea* occupies this position in Figure 7a, and all other analyses are indecisive on this point. Eudicots (angiosperms with triaperturate or triaperturate-derived pollen; here, *Platanus, Caltha, Trochodendron, Dillenia, Hamamelis, Chrysolepis, Betula, Casuarina*) are monophyletic in the combined nucleotide and string analyses (Figs. 6a, b, 7a, b) (see Chase et al., 1993) but are polyphyletic in the combined amino acid analyses (Fig. 8a, b). For further discussion and reference to cladograms based solely on the morphological evidence, see Nixon et al. (1994).

The topological differences resulting from use of either *rbc*L nucleotide, string, or amino acid data might imply that different sets of morphological characters (of Nixon et al., 1994) show congruence with these different data forms. If one were to hold the evidential significance of the morphological data constant, one might identify those portions of primary *rbc*L nucleotide sequence that were incongruent under each data form and ignore them in future studies. Alternatively, one could take the opposite approach and ignore those Nixon et al. (1994) characters that were not congruent among all *rbc*L data forms. We suggest that either approach is nihilistic with respect to either *rbc*L or morphology; because congruence is an aspect of total interaction, the utility of either set of evidence is always judged relative to the other. Nevertheless, hierarchic correlation can be directed at one subset of total evidence if, as in the case of *rbc*L, it is reasonable to assume a single, unifying functional history. If an investigator were willing to hold all evidence except *rbc*L constant, hypotheses of correlation between functional constraints

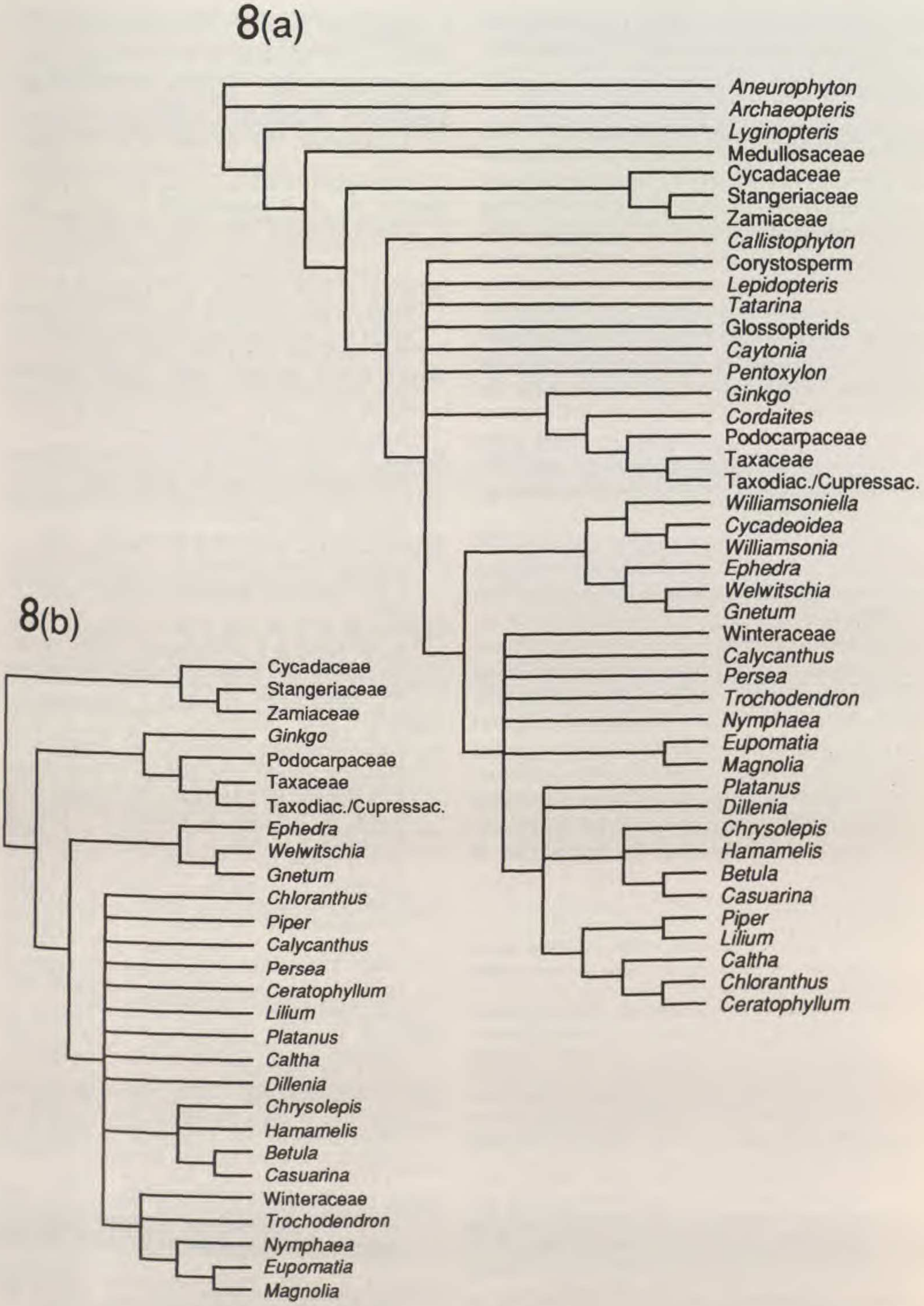


FIGURES 6–8. Total evidence analyses of morphological and *rbc*L data for fossil and extant seed plants. The morphological data and taxon sampling of Nixon et al. (1994; matrix version as of 8 November 1993) was followed for cladistic analyses of fossil and living seed plants (the "a" series) and of extant seed plants only (the "b" series). For both taxonomic scopes, *rbc*L evidence was combined as one of three data forms: nucleotide sequences (6), nucleotide string recognitions (7), or amino acid sequences (8) obtained from single organisms (see Table 2). For

→



analyses including fossil taxa, *rbc*L character states were scored as missing (i.e., "?"; cf. Platnick et al., 1991; Swofford, 1993: 21–24). Topological results (from PAUP 3.1.1; Swofford, 1993) shown represent either single trees or the strict consensus (= combinable component consensus in all cases) of all most-parsimonious trees found (see Table 5). See text for further discussion.


8(a)
Aneurophyton
Archaeopteris
Lyginopteris
Medullosaceae
Cycadaceae
Stangeriaceae
Zamiaceae
Callistophyton
Corystosperm
Lepidopteris
Tatarina
Glossopterids
Caytonia
Pentoxylon
Ginkgo
Cordaites
Podocarpaceae
Taxaceae
Taxodiac./Cupressac.
Williamsoniella
Cycadeoidea
Williamsonia
Ephedra
Welwitschia
Gnetum
Winteraceae
Calycanthus
Persea
Trochodendron
Nymphaea
Eupomatia
Magnolia
Platanus
Dillenia
Chrysolepis
Hamamelis
Betula
Casuarina
Piper
Lilium
Caltha
Chloranthus
Ceratophyllum
8(b)
Cycadaceae
Stangeriaceae
Zamiaceae
Ginkgo
Podocarpaceae
Taxaceae
Taxodiac./Cupressac.
Ephedra
Welwitschia
Gnetum
Chloranthus
Piper
Calycanthus
Persea
Ceratophyllum
Lilium
Platanus
Caltha
Dillenia
Chrysolepis
Hamamelis
Betula
Casuarina
Winteraceae
Trochodendron
Nymphaea
Eupomatia
Magnolia

and phylogenetic history could be generated from the congruence patterns of each *rbc*L character.

Conclusions

The phylogenetic informativeness of *rbc*L variation is obviously subject to any special properties the gene may have. Unlike for most morphological characters, some such properties can be listed for *rbc*L with confidence: (i) *rbc*L nucleotides show clocklike substitutional behavior, which may either help or hinder tree reconstruction depending upon the temporal depth and asymmetry of a given phylogenetic question; (ii) strong functional constraints exist over the majority of informative nucleotide characters, which is expected from (i) under the neutral theory; and (iii) the form that *rbc*L evidence takes (e.g., nucleotides, strings, or amino acids) does not appreciably affect its interaction with other evidence containing diverse functional histories (e.g., morphological data).

Although *rbc*L trees often appear consistent with taxonomic opinion (or are substantially congruent with other cladistic topologies), their power as lone cladistic tools will always be restricted by the intrinsic limits of internal evaluation of data. Because *rbc*L sequences clearly have a unifying functional history, simultaneous study of *all* available evidence become imperative. Functional constraints on *rbc*L, rDNA, or endosperm evolution are not expected to be similar; therefore patterns of character congruence among such diverse information sources will provide hypotheses of cladogenetic history significantly more powerful than studies of *rbc*L alone.

Literature Cited

Albert, V. A. & B. D. Mishler. 1992. On the rationale and utility of weighting nucleotide sequence data. Cladistics 8: 73–83.

———, A. Backlund & K. Bremer. DNA characters and cladistics: The optimization of functional history. *In* Models in Phylogenetic Reconstruction. The Systematics Association, Oxford Univ. Press. (in press).

———, M. W. Chase & B. D. Mishler. 1993. Character-state weighting for cladistic analysis of protein-coding DNA sequences. Ann. Missouri Bot. Gard. 80: 752–766.

———, B. D. Mishler & M. W. Chase. 1992a. Character-state weighting for restriction site data in phylogenetic reconstruction, with an example from chloroplast DNA. Pp. 369–403 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

———, K. Bremer, M. W. Chase, J. Manhart, B. D. Mishler & K. C. Nixon. 1992b. *rbc*L gene sequences and phylogenetic studies of vascular plants. Pp. 13–14 *in* program booklet, "Origin and Relationships of the Major Plant Groups," 39th Annual Systematics Symposium, Missouri Bot. Gard., St. Louis [abstract].

———, S. E. Williams & M. W. Chase. 1992c. Carnivorous plants: Phylogeny and structural evolution. Science 257: 1491–1495.

Andersson, I., S. Knight, G. Schneider, Y. Lindqvist, T. Lundqvist, C.-I. Brånden & G. H. Lorimer. 1989. Crystal structure of the active site of ribulose-bisphosphate carboxylase. Nature 337: 229–234.

Barrett, M., M. J. Donoghue & E. Sober. 1991. Against consensus. Syst. Zool. 40: 486–493.

———, ——— & ———. 1993. Crusade? A reply to Nelson. Syst. Biol. 42: 216–217.

Bousquet, J., S. H. Strauss, A. D. Doerksen & R. A. Price. 1992. Extensive variation in evolutionary rate of *rbc*L gene sequences among seed plants. Proc. Natl. Acad. Sci. U.S.A. 89: 7844–7848.

Bremer, B. & K. Bremer. 1989. Cladistic analysis of blue-green procaryote interrelationships and chloroplast origin based on 16S rRNA oligonucleotide catalogues. J. Evol. Biol. 2: 13–30.

Bremer, K. 1985. Summary of green plant phylogeny and classification. Cladistics 1: 369–385.

———. 1990. Combinable component consensus. Cladistics 6: 369–372.

Bruneau, A., J. J. Doyle & J. D. Palmer. 1990. A chloroplast DNA inversion as a subtribal character in the Phaseolae (Leguminosae). Syst. Bot. 15: 378–386.

Chapman, M. S., S. W. Suh, P. M. G. Curmi, D. Cascio, W. W. Smith & D. S. Eisenberg. 1988. Tertiary structure of plant RuBisCo: Domains and their contacts. Science 241: 71–74.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qiu, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedrén, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. S. Soltis, S. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, Jr., S. W. Graham, S. C. H. Barrett, S. Dayanandan & V. A. Albert. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Missouri Bot. Gard. 80: 528–580.

Clegg, M. T. 1993. Chloroplast gene sequences and the study of plant evolution. Proc. Natl. Acad. Sci. U.S.A. 90: 363–367.

——— & G. Zurawski. 1992. Chloroplast DNA and the study of plant phylogeny: Present status and future prospects. Pp. 1–13 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

Conti, E., A Fischbach & K. J. Sytsma. 1993. Tribal relationships in Onagraceae: Implications from *rbc*L sequence data. Ann. Missouri Bot. Gard. 80: 672–685.

Crane, P. R. 1985. Phylogenetic analysis of seed plants and the origin of angiosperms. Ann. Missouri Bot. Gard. 72: 716–793.

Dayhoff, M. O., R. M. Schwartz & B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. Dayhoff (editor), Atlas of

Protein Sequence and Structure, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

DeBry, R. W. & N. A. Slade. 1985. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. Syst. Zool. 34: 24–34.

De Queiroz, A. 1993. For consensus (sometimes). Syst. Biol. 42: 368–372.

Donoghue, M. J. 1989. Phylogenies and the analysis of evolutionary sequences, with examples from seed plants. Evolution 43: 1137–1156.

——— & J. A. Doyle. 1989. Phylogenetic analysis of angiosperms and the relationships of Hamamelidae. Pp. 17–45 *in* P. R. Crane & S. Blackmore (editors), Evolution, Systematics, and Fossil History of the Hamamelidae, Volume 1: Introduction and "Lower" Hamamelidae. Clarendon Press, Oxford.

——— & M. J. Sanderson. 1992. The suitability of molecular and morphological evidence in reconstructing plant phylogeny. Pp. 340–368 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

Downie, S. R. & J. D. Palmer. 1992. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. Pp. 14–35 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

———, R. G. Olmstead, G. Zurawski, D. E. Soltis, P. S. Soltis, J. C. Watson & J. D. Palmer. 1991. Six independent losses of the chloroplast DNA *rpl*2 intron in dicotyledons: Molecular and phylogenetic implications. Evolution 45: 1245–1259.

Doyle, J. A. & M. J. Donoghue. 1986. Seed plant phylogeny and the origin of angiosperms: An experimental cladistic approach. Bot. Rev. 52: 321–431.

——— & ———. 1992. Fossils and seed plant phylogeny reanalyzed. Brittonia 44: 89–106.

Doyle, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17: 144–163.

Farris, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. 18: 374–385.

———. 1970. Methods for computing Wagner trees. Syst. Zool. 19: 83–92.

———. 1978. Inferring phylogenetic trees from chromosome inversion data. Syst. Zool. 27: 275–284.

———. 1979. The information content of the phylogenetic system. Syst. Zool. 28: 483–519.

———. 1983. The logical basis of phylogenetic analysis. Pp. 7–36 *in* N. I. Platnick & V. A. Funk (editors), Advances in Cladistics, Vol. 2. Columbia Univ. Press, New York.

———. 1989a. The retention index and the rescaled consistency index. Cladistics 5: 417–419.

———. 1989b. The retention index and homoplasy excess. Syst. Zool. 38: 406–407.

———. 1991. Excess homoplasy ratios. Cladistics 7: 81–91.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27: 401–410.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19: 99–113.

———. 1971. Toward defining the course of evolution: Minimum change for specific tree topology. Syst. Zool. 20: 406–416.

———. 1986. The estimate of total nucleotide substitutions from pairwise differences is biased. Philos. Trans., Ser. B 316: 317–324.

——— & E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4: 579–593.

Fox, G. E., E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen & C. R. Woese. 1980. The phylogeny of prokaryotes. Science 209: 457–463.

Gaut, B. S., S. V. Muse, W. D. Clark & M. T. Clegg. 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. J. Molec. Evol. 35: 292–303.

Golenberg, E. M., D. E. Giannasi, M. T. Clegg, C. J. Smiley, M. Durbin, D. Henderson & G. Zurawski. 1990. Chloroplast DNA sequence from a Miocene *Magnolia* species. Nature 334: 656–658.

Goloboff, P. A. 1991. Homoplasy and the choice among cladograms. Cladistics 7: 215–232.

Hamby, R. K. & E. A. Zimmer. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. Pp. 50–91 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

Hendy, M. D. & D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38: 297–309.

Hudson, G. S., J. D. Mahon, P. A. Anderson, M. J. Gibbs, M. R. Badger, T. J. Andrews & P. R. Whitfeld. 1990. Comparisons for *rbc*L genes for the large subunit of ribulose-bisphosphate carboxylase from closely related C_3 and C_4 plant species. J. Biol. Chem. 265: 808–814.

Jansen, R. K. & J. D. Palmer. 1987. Chloroplast DNA from lettuce and *Barnadesia* (Asteraceae): Structure, gene localization, and characterization of a large inversion. Curr. Genet. 11: 553–564.

Jones, T. R., A. G. Kluge & A. J. Wolf. 1993. When theories and methodologies clash: A phylogenetic reanalysis of the North American ambystomatid salamanders. Syst. Biol. 42: 92–102.

Jukes, T. H. & C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro (editor), Mammalian Protein Metabolism. Academic Press, New York.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Molec. Evol. 16: 111–120.

———. 1983. The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, Cambridge.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38: 7–25.

——— & J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18: 1–32.

Kron, K. A. & M. W. Chase. 1993. Systematics of the Ericaceae, Empetraceae, Epacridaceae and related taxa based upon *rbc*L sequence data. Ann. Missouri Bot. Gard. 80: 735–741.

LAVIN, M. & M. LUCKOW. 1993. Origins and relationships of tropical North America in the context of the boreotropics hypothesis. Amer. J. Bot. 80: 1–14.

———, J. J. DOYLE & J. D. PALMER. 1990. Evolutionary significance of the loss of chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. Evolution 44: 390–402.

LES, D. H. 1988. The origin and affinities of the Ceratophyllaceae. Taxon 37: 326–345.

———, D. K. GARVIN & C. F. WIMPEE. 1991. Molecular evolutionary history of ancient aquatic angiosperms. Proc. Natl. Acad. Sci. U.S.A. 88: 10119–10123.

LOCONTE, H. & D. W. STEVENSON. 1990. Cladistics of the Spermatophyta. Brittonia 42: 197–211.

——— & ———. 1991. Cladistics of the Magnoliidae. Cladistics 7: 267–296.

MANHART, J. R. Phylogenetic analysis of green plant *rbc*L sequences. Molecular Phylogenetics and Evolution (in press).

MISHLER, B. D. 1994. Cladistic analysis of molecular and morphological data. Amer. J. Phys. Anthropol. 94: 143–156.

——— & S. P. Churchill. 1985. Transition to a land flora: Phylogenetic relationships of the green algae and bryophytes. Cladistics 1: 305–328.

———, L. A. LEWIS, M. A. BUCHHEIM, K. S. RENZAGLIA, D. J. GARBARY, C. F. DELWICHE, F. W. ZECHMAN, T. S. KANTZ & R. L. CHAPMAN. 1994. Phylogenetic relationships of the "green algae" and "bryophytes." Ann. Missouri Bot. Gard. 81: 451–483.

NEI, M. 1987. Molecular Evolutionary Genetics. Columbia Univ. Press, New York.

NELSON, G. 1993. Why crusade against consensus? A reply to Barrett, Donoghue, and Sober. Syst. Biol. 42: 215–216.

NIXON, K. C., W. L. CREPET, D. STEVENSON & E. M. FRIIS. 1994. A reevaluation of seed plant phylogeny. Ann. Missouri Bot. Gard. 81: 484–533.

NORELL, M. A. & M. J. NOVACEK. 1992. Congruence between superpositional and phylogenetic patterns: Comparing cladistic patterns with fossil records. Cladistics 8: 319–337.

OLMSTEAD, R. G. 1989. Phylogeny, phenotypic evolution, and biogeography of the *Scutellaria angustifolia* complex (Lamiaceae); inference from morphological and molecular data. Syst. Bot. 14: 320–338.

PAGE, R. D. M. 1993. Genes, organisms, and areas: The problem of multiple lineages. Syst. Biol. 42: 77–84.

PALMER, J. D., B. OSORIO & W. F. THOMPSON. 1988. Evolutionary significance of inversion in legume chloroplast DNAs. Curr. Genet. 14: 65–74.

PENNY, D., M. D. HENDY & M. A. STEEL. 1991. Testing the theory of descent. Pp. 155–193 *in* M. M. Miyamoto & J. Cracraft (editors), Phylogenetic Analysis of DNA Sequences. Oxford Univ. Press, New York.

PERUTZ, M. F. & H. LEHMAN. 1968. Molecular pathology of human haemoglobin. Nature 219: 902–909.

PLATNICK, N. I., C. E. GRISWOLD & J. A. CODDINGTON. 1991. On missing entries in cladistic analysis. Cladistics 7: 337–343.

QIU, Y.-L., M. W. CHASE, D. H. LES, H. G. HILLS & C. R. PARKS. 1993. Molecular phylogenetics of the Magnoliidae: Cladistic analyses of nucleotide sequences of the plastid gene *rbc*L. Ann. Missouri Bot. Gard. 80: 587–606.

RAUBESON, L. A. & R. K. JANSEN. 1992. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255: 1697–1699.

RIEDL, R. 1978. Order in Living Organisms. Wiley, New York.

RODMAN, J., R. A. PRICE, K. G. KAROL, E. CONTI, K. J. SYTSMA & J. D. PALMER. 1993. Nucleotide sequences of the *rbc*L gene indicate monophyly of mustard oil plants. Ann. Missouri Bot. Gard. 80: 686–699.

SOLTIS, P. S., D. E. SOLTIS & C. J. SMILEY. 1992. An *rbc*L sequence from a Miocene *Taxodium* (bald cypress). Proc. Natl. Acad. Sci. U.S.A. 89: 449–451.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1.1. Computer program and user's manual distributed by the Illinois Natural History Survey, Champaign, Illinois.

——— & W. P. MADDISON. 1987. Reconstructing ancestral states under Wagner parsimony. Math. Biosci. 87: 199–229.

SYTSMA, K. J., J. F. SMITH & P. E. BERRY. 1991. Biogeography and evolution of morphology, breeding systems, flavonoids, and chloroplast DNA in the four Old World species of *Fuchsia* (Onagraceae). Syst. Bot. 16: 257–269.

TAYLOR, D. W. & L. J. HICKEY. 1992. Phylogenetic evidence for the herbaceous origin of angiosperms. Pl. Syst. Evol. 180: 137–156.

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage maps with particular reference to the evolution of humans and the apes. Evolution 37: 221–244.

TROITSKY, A. V., Y. F. MELEKHOVETS, G. M. RAKHIMOVA, V. K. BOBROVA, K. M. VALIEJO-ROMAN & A. S. ANTONOV. 1991. Angiosperm origin and early stages of seed plant evolution deduced from rRNA sequence comparisons. J. Molec. Evol. 32: 253–261.

WENDEL, J. F. & V. A. ALBERT. 1992. Phylogenetics of the cotton genus (Gossypium): Character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. Syst. Bot. 17: 115–143.

WILSON, M. A., B. GAUT & M. T. CLEGG. 1990. Chloroplast DNA evolves slowly in the palm family. Molec. Biol. Evol. 7: 303–314.

ZIMMER, E. A., R. K. HAMBY, M. L. ARNOLD, D. A. LEBLANC & E. C. THERIOT. 1989. Ribosomal RNA phylogenies and flowering plant evolution. Pp. 205–214 *in* B. Fernholm, K. Bremer & H. Jörnvall (editors), The Hierarchy of Life: Molecules and Morphology in Phylogenetic Analysis. Elsevier Science Publishers, Amsterdam.

APPENDIX 1 (pp. 554–562).* Inferred amino acid changes on the internal branches of a nucleotide-based cladogram (one of eight equally most-parsimonious).

This table and accompanying cladogram contain information about the functional impact of specific nucleotide changes (as reflected by alterations in amino acid identity). Following the apomorphy list format of PAUP 3.1.1 (Swofford, 1993), each internal branch of the reference tree is identified by the nodes it connects. For each node pair, optimized nucleotide changes are identified by position ("POS," i.e., the 1–1428 bases of the *rbc*L gene used), character consistency index ("*c*," each of which represents a separate contribution of the ensemble consistency of the entire tree; see Farris 1989a), the actual change inferred ("NUCΔ," with arrows following the conventions in the PAUP 3.1.1 manual; Swofford, 1993: 121), amino acid changes ("AA") that occur at this position (listed nondirectionally; see below), and their substitutional category ("SC") as determined from the PAM-250 log-odds matrix of Dayhoff et al. (1978: 352; log-odds scores of 0 and above are considered labile (L), whereas negative values are here considered nonlabile (NL); potentially nonlabile (PNL) indicates mixed-odds changes at the codon involving a given position, and synonymous changes (constant amino acid identity) are indicated by "—").

For example, a line of the following form

175 1.00 c → g R, L, A NL

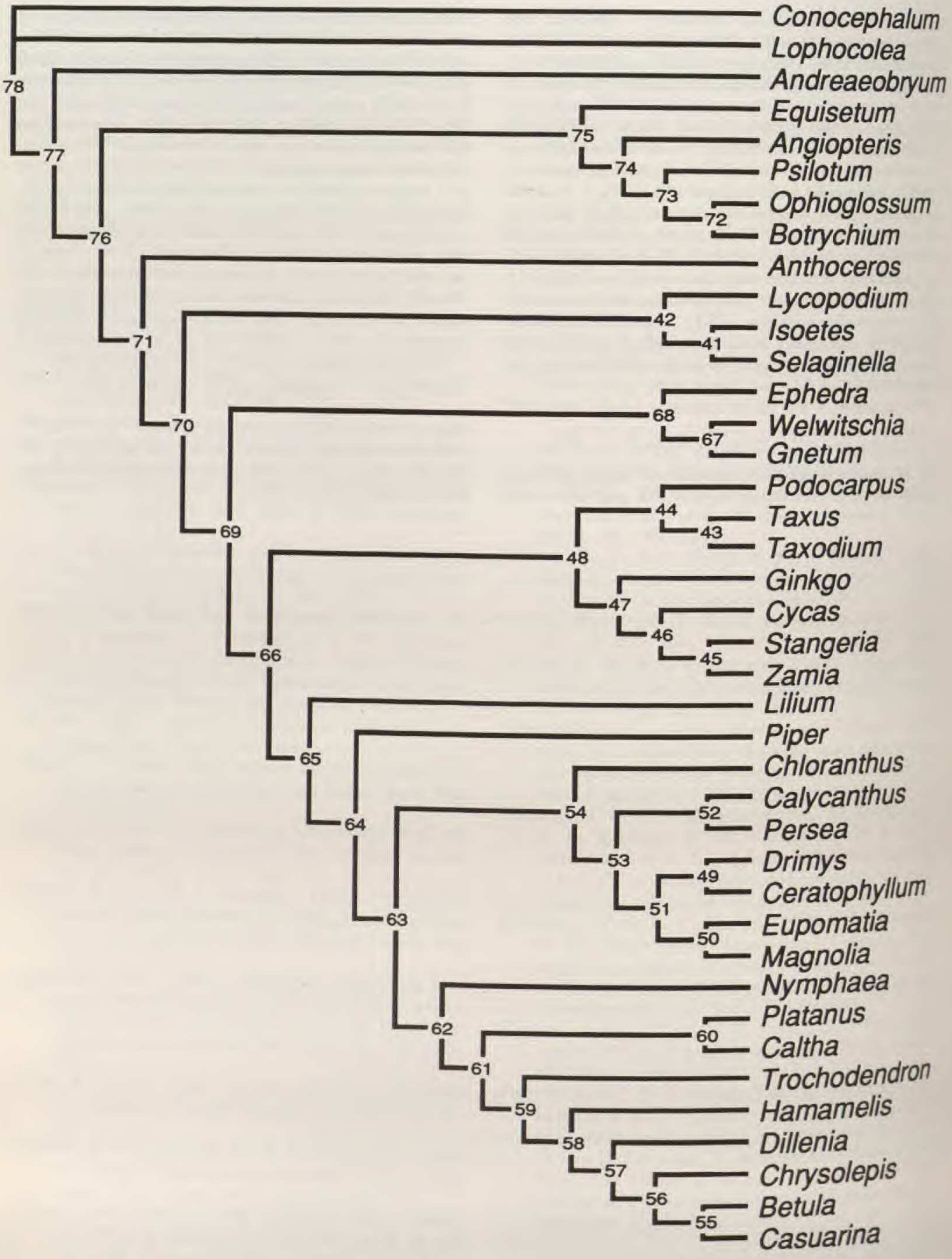
can be readily diagnosed: character 175 changes from nucleotide C to nucleotide G (on this particular tree; constancy of character-state reconstruction among all 8 trees would be indicated by a double-lined arrow) with a *c* of 1.000 (i.e., no homoplasy), and the codon in which character 175 belongs changes between the amino acids R, L, and A (using standard IUB amino acid codes; see Nei, 1987: 24; Swofford, 1993: 67). Note, however, that this does not necessarily mean that this particular character-state change gives the indicated changes in amino acid sequence; rather, it merely indicates that it *might* be involved in the changes (i.e., the C → G nucleotide transformation may not affect amino acid identify at all; thus, the indicated amino acid changes are the "worst" that can happen under the influence of character 175). The NL designation indicates that any pairwise transformation between R, L, and A would represent a nonlabile change.

In the line below

486 0.167 a → g L, S —

there is a nucleotide transformation in position 486, yet it can be positively diagnosed as *not* responsible for the different amino acid identities in its associated codon (thus, the SC is given as "—").

* Correction added in proof: P. 560, under "NODE 62–61," third line from bottom, right hand column, should read "L."


Conocephalum
Lophocolea
Andreaeobryum
Equisetum
Angiopteris
Psilotum
Ophioglossum
Botrychium
Anthoceros
Lycopodium
Isoetes
Selaginella
Ephedra
Welwitschia
Gnetum
Podocarpus
Taxus
Taxodium
Ginkgo
Cycas
Stangeria
Zamia
Lilium
Piper
Chloranthus
Calycanthus
Persea
Drimys
Ceratophyllum
Eupomatia
Magnolia
Nymphaea
Platanus
Caltha
Trochodendron
Hamamelis
Dillenia
Chrysolepis
Betula
Casuarina
78
77
76
75
74
73
72
71
70
42
41
69
68
67
66
48
44
43
47
46
45
65
64
63
54
52
53
51
49
50
62
61
60
59
58
57
56
55

NODE 78-77

POS	*c*	NUCΔ	AA	SC
68	0.200	a->c	T,N	L
69	0.500	c->t	T,N	-
102	0.429	a->t	constant	-
150	0.231	a->t	A,P,S	L
165	0.231	a->t	A,W	NL
175	1.000	c->g	R,L,A	NL
186	0.222	a->c	constant	-
204	0.375	a->t	constant	-
342	1.000	a=>c	constant	-
345	0.333	t->c	constant	-
391	0.333	a=>c	R,P	L
405	0.222	t->a	constant	-
433	0.250	a->t	T,V,S,I	PNL
435	0.300	a->c	T,V,S,I	PNL
552	0.200	t=>c	constant	-
696	0.286	t=>a	constant	-
711	0.250	a->g	constant	-
740	0.667	c=>g	S,C,Y	L
764	0.400	c=>a	A,Q,E,V,H,I	PNL
767	0.333	g=>t	C,F	NL
783	0.600	t->a	constant	-
785	0.200	t->c	V,M,A	PNL
786	0.500	a->t	V,M,A	L
789	0.429	a=>t	constant	-
810	0.333	a->g	constant	-
840	0.167	a=>g	L,S	-
844	0.200	c->t	H,Y,S,F	PNL
906	0.286	a->c	D,R	-
958	0.500	t->a	L,M	L
1027	0.167	c->t	constant	-
1035	0.250	c->t	constant	-
1038	0.500	a->t	constant	-
1072	0.500	a=>c	constant	-
1095	0.500	a=>t	constant	-
1101	0.250	t->c	constant	-
1134	1.000	a->t	constant	-
1212	0.429	t=>a	constant	-
1227	0.286	t=>c	H,Q	-
1260	0.250	c->t	constant	-
1290	1.000	a=>t	constant	-
1345	0.154	g->a	A,S,T,C	L
1346	0.125	c=>g	A,S,T,C	L

NODE 77-76

POS	*c*	NUCΔ	AA	SC
120	0.167	t->c	constant	-
357	0.286	t->c	constant	-
426	1.000	a=>t	V,P,L,T,I	NL
567	1.000	a=>t	constant	-
660	0.167	t->c	constant	-
816	0.750	t->a	constant	-
852	0.286	c->t	constant	-
876	0.143	c->t	constant	-
927	0.231	g->t	I,M	L
930	0.111	t=>c	constant	-
960	1.000	a=>g	L,M	-
963	0.182	t->c	C,S	-
981	0.143	c->t	constant	-
984	0.182	t->c	A,S,T	-
1018	0.250	c->g	Q,E,D	L
1107	0.333	t=>a	constant	-
1111	0.286	t=>a	L,M,T	L
1113	0.500	a=>g	L,M,T	-
1116	0.222	a->t	P,A	-
1125	0.375	a->g	L,S,F,I,M	L
1137	0.231	a->g	constant	-
1176	0.250	a->g	E,D	-
1254	0.429	t=>a	constant	-
1363	0.167	c->t	constant	-

NODE 76-71

POS	*c*	NUCΔ	AA	SC
60	0.250	t->c	constant	-
111	0.200	a=>g	constant	-
138	0.273	t->a	P,L	-
165	0.231	t->c	A,W	-
225	0.333	t=>c	constant	-
258	0.333	t=>c	G,D,E,N,H	-
321	0.333	a=>g	constant	-
327	0.167	a->g	constant	-
351	0.167	t=>c	constant	-
486	0.167	a->g	L,S	-
564	0.214	t->a	A,V	-
603	0.143	a->g	constant	-
615	0.125	t->c	constant	-
682	0.333	t->g	S,A	L
708	0.200	a=>g	constant	-
759	0.333	a=>g	constant	-
785	0.200	c->t	V,M,A	PNL
786	0.500	t->g	V,M,A	L
844	0.200	t->c	H,Y,S,F	PNL
858	0.167	t=>c	constant	-
1021	0.333	g->a	V,I,L,M	L
1062	0.500	c=>t	I,Y	-
1198	0.167	t->c	L,S	-
1212	0.429	a->g	constant	-
1320	0.143	a=>g	Q,E,A	-
1335	0.167	t->c	constant	-
1398	0.250	a->g	R,K,I	-

NODE 71-70

POS	*c*	NUCΔ	AA	SC
88	0.143	g->a	E,K,Q,T	L
138	0.273	a->c	P,L	-
150	0.231	t=>c	A,P,S	-
153	0.111	a=>g	constant	-
261	0.167	t->c	I,L	-
315	0.167	a->g	constant	-
342	1.000	c=>t	constant	-
387	0.333	t=>c	constant	-
405	0.222	a->g	constant	-
414	0.167	a=>g	S,L	-
444	0.167	t->c	F,C	-
510	0.167	a->g	constant	-
519	0.182	c->t	constant	-
711	0.250	g->a	constant	-
720	0.200	a->g	constant	-
792	0.500	t->c	I,S	-
795	0.250	t->c	V,G	-
822	0.143	t->c	constant	-
876	0.143	t->c	constant	-
981	0.143	t->c	constant	-
1071	0.167	t=>c	constant	-
1101	0.250	c->t	constant	-
1128	0.222	t->c	constant	-
1149	0.111	t->c	constant	-
1168	0.250	t=>c	L,*	-
1170	0.143	a=>g	L,*	-
1179	0.400	t=>c	constant	-
1245	0.200	t->a	constant	-
1363	0.167	t->c	constant	-

NODE 70-42

POS	*c*	NUCΔ	AA	SC
48	0.500	t=>c	constant	-
102	0.429	t=>c	constant	-
109	0.333	t->c	constant	-
280	0.250	g->a	D,E,K,T	L
313	0.250	t->c	constant	-
336	0.250	t->c	constant	-
403	0.333	c=>t	constant	-
447	0.167	a->g	Q,M,I,T,L,W	L
453	0.273	a->g	constant	-
543	0.333	t->c	constant	-
597	0.167	t->c	constant	-
630	0.300	a->g	P,A	-
648	0.200	t->c	constant	-
690	0.429	t=>c	A,G,T	-
699	0.250	t->c	constant	-
711	0.250	a->t	constant	-
729	0.250	t->c	constant	-
840	0.167	g->a	L,S	-
843	1.000	t=>c	A,S	-
870	0.600	t->a	constant	-
885	0.286	t->c	constant	-
915	0.167	a->g	K,R	-
975	0.333	t=>c	constant	-
982	0.182	g->t	A,S,T	L
1005	0.375	t->c	constant	-
1018	0.250	g->c	Q,E,D	L
1042	0.167	t->c	L,S	-
1068	0.333	a->g	K,R,E	-
1077	0.200	t->c	constant	-
1176	0.250	g->a	E,D	-
1198	0.167	c->t	L,S	-
1206	0.111	t->c	constant	-
1221	0.200	a=>g	L,S	-
1236	0.333	a=>g	constant	-
1260	0.250	t->c	constant	-
1329	0.600	a=>g	D,E	-
1335	0.167	c->t	constant	-
1345	0.154	a=>t	A,S,T,C	L
1350	0.250	a->g	constant	-
1371	0.200	t->c	constant	-

NODE 42-41

POS	*c*	NUCΔ	AA	SC
132	0.286	t->c	constant	-
189	0.375	t=>c	constant	-
207	0.600	t=>c	constant	-
225	0.333	c=>t	constant	-
267	0.375	t=>c	P,T	-
297	0.200	t=>c	A,V,C	-
324	0.167	t=>c	constant	-
441	0.286	t=>c	constant	-
459	0.250	t=>c	constant	-
528	0.429	a->g	constant	-
567	1.000	t->c	constant	-
676	0.500	t=>a	Y,N,F	NL
696	0.286	a=>g	constant	-
702	0.200	a=>g	constant	-
718	0.333	t=>c	constant	-
744	0.250	a=>g	constant	-
768	0.167	t=>c	C,F	-
780	0.143	a=>g	constant	-
855	0.400	a=>g	constant	-
897	0.667	a=>t	A,V	-
912	0.333	a=>g	constant	-
969	0.429	a=>g	constant	-
993	0.750	a=>g	constant	-

996	0.400	a=>g	constant	-
1011	0.429	a=>g	constant	-
1021	0.333	a->g	V,I,L,M	L
1095	0.500	t=>c	constant	-
1137	0.231	g->a	constant	-
1140	0.300	a=>g	constant	-
1164	0.333	t=>c	P,Q,S	-
1212	0.429	g->c	constant	-
1296	0.333	t=>c	constant	-
1377	1.000	t=>c	constant	-

NODE 70-69

POS	c	NUCΔ	AA	SC
10	0.500	c=>a	- excluded, (PRIMER)	
15	1.000	g=>a	- excluded, (PRIMER)	
75	0.125	c->t	Y,F	-
84	0.214	t=>g	D,E,Q	L
88	0.143	a->c	E,K,Q,T	L
108	0.400	t=>c	I,T	-
124	1.000	a=>g	M,V,L	L
126	1.000	g=>a	M,V,L	-
165	0.231	c->a	A,W	-
201	0.250	t=>c	constant	-
246	0.333	t=>a	constant	-
271	0.250	g->c	P,A,V,T	L
318	0.250	t=>c	constant	-
321	0.333	g=>t	constant	-
327	0.167	g->a	constant	-
388	0.333	t=>c	constant	-
397	0.333	t->c	L,S,I	-
486	0.167	g->a	L,S	-
504	0.167	t->c	constant	-
522	0.286	t->c	constant	-
660	0.167	c->t	constant	-
661	1.000	g=>t	V,C	-
662	1.000	t=>g	V,C	-
663	0.500	a->c	V,C	-
672	0.300	t->a	constant	-
673	0.111	c=>a	L,I	-
764	0.400	a=>t	A,Q,E,V,H,I	NL
786	0.500	g->t	V,M,A	-
810	0.333	g->a	constant	-
837	0.300	t->c	S,I,T	-
852	0.286	t->c	constant	-
864	0.333	t=>c	constant	-
906	0.286	c->t	D,R	-
927	0.231	t->g	I,M	L
940	0.250	t=>c	L,S	-
1017	0.333	t=>a	constant	-
1023	0.231	a=>c	V,I,L,M	-
1058	0.500	a=>t	Y,F,C,L	L
1116	0.222	t=>a	P,A	-
1123	0.250	t=>c	L,S,F,I,M	L
1212	0.429	g->a	constant	-
1330	0.167	a->g	I,V	L
1389	0.143	a->g	constant	-
1392	0.143	a->g	constant	-

NODE 69-66

POS	c	NUCΔ	AA	SC
90	0.250	g->a	E,K,Q,T	-
144	0.333	a->g	constant	-
177	0.300	t=>c	R,L,A	-
264	0.333	a->g	D,E	-
267	0.375	t=>c	P,T	-
276	0.286	a->g	constant	-
279	0.167	a=>g	constant	-
315	0.167	g->a	constant	-
412	0.200	t=>c	S,L	-
505	0.200	t->c	constant	-
534	0.200	a=>g	constant	-
549	0.200	a->g	constant	-
600	0.333	t=>c	constant	-
624	0.667	t=>c	constant	-
663	0.500	c->t	V,C	-
687	0.167	a->g	constant	-
696	0.286	a=>g	constant	-
780	0.143	a=>g	constant	-
813	0.231	a=>g	constant	-
861	0.143	t->c	constant	-
963	0.182	c->t	C,S	-
1029	0.250	a->g	constant	-
1047	0.429	t=>g	constant	-
1056	0.167	c->t	constant	-
1140	0.300	a=>g	constant	-
1173	0.167	t=>c	constant	-
1185	0.200	a=>g	constant	-
1203	0.200	a=>g	constant	-
1398	0.250	g->a	R,K,I	-

NODE 66-48

POS	c	NUCΔ	AA	SC
33	0.500	t=>c	V,S,F,D,A	-
84	0.214	g=>a	D,E,Q	-
138	0.273	c->t	P,L	-
243	0.200	a=>g	constant	-
290	0.125	a->t	Y,F	L
297	0.200	t=>c	A,V,C	-
309	0.143	t->c	constant	-
312	0.182	t->c	P,F	-
346	0.333	a->c	M,L	L
498	0.333	t->c	constant	-
546	0.250	t=>c	constant	-
552	0.200	c->t	constant	-
570	0.200	t=>c	constant	-
612	0.111	a->g	constant	-
639	0.333	t=>c	constant	-
656	0.500	t->g	L,V,C	-
657	1.000	a=>c	L,V,C	NL
693	0.167	a->g	constant	-
771	0.375	t->c	constant	-
808	0.167	t=>c	constant	-
810	0.333	a->g	constant	-
822	0.143	c->t	constant	-
885	0.286	t=>c	constant	-
914	0.143	a->g	K,R	L
954	0.286	a=>g	L,S	-
1021	0.333	a->g	V,I,L,M	L
1042	0.167	t=>c	L,S	-
1221	0.200	a=>g	L,S	-
1245	0.200	a->t	constant	-
1320	0.143	g=>a	Q,E,A	-
1332	0.500	t=>g	I,V	-
1359	0.286	t->c	P,A,L	-
1416	0.667	g->t	I,M,V,W	L
1422	0.429	g=>t	T,V,L,K	-

NODE 48-44

POS	c	NUCΔ	AA	SC
90	0.250	a->g	E,K,Q,T	-
147	0.154	a=>c	constant	-
264	0.333	g->a	D,E	-
276	0.286	g->a	constant	-
393	0.231	a->g	R,P	-
456	0.222	t->a	constant	-
471	0.500	t=>g	A,V	-
505	0.200	c->t	constant	-
538	0.400	t=>c	L,I	-
718	0.333	t->c	constant	-
759	0.333	g=>a	constant	-
768	0.167	t=>c	C,F	-
825	0.375	t=>g	T,I	-
835	0.500	a=>t	S,I,T	L
836	0.222	g=>c	S,I,T	L
837	0.300	c=>g	S,I,T	-
1023	0.231	c->a	V,I,L,M	-
1122	0.400	t=>c	constant	-
1128	0.222	c->t	constant	-
1198	0.167	c->t	L,S	-
1224	0.429	a->g	constant	-
1263	0.500	a=>g	R,*	-
1389	0.143	g->a	constant	-
1397	1.000	a->t	R,K,I	NL
1413	1.000	a->t	T,A,S,E,P	-

NODE 44-43

POS	c	NUCΔ	AA	SC
10	0.500	a->c	- excluded,	
18	0.333	g->a	- excluded,	
81	0.333	t->a	constant	-
258	0.333	c=>t	G,D,E,N,H	-
284	0.286	a=>g	N,D,S,T,E,G	L
318	0.250	c=>t	constant	-
414	0.167	g=>a	S,L	-
435	0.300	c=>a	T,V,S,I	-
450	0.214	t=>c	constant	-
498	0.333	c->t	constant	-
504	0.167	c=>t	constant	-
507	0.333	a=>g	constant	-
522	0.286	c->a	constant	-
564	0.214	a->c	A,V	-
579	0.375	t=>c	constant	-
612	0.111	g->a	constant	-
618	0.333	a=>g	constant	-
702	0.200	a=>g	constant	-
813	0.231	g->a	constant	-
952	0.500	t=>c	L,S	-
984	0.182	c->t	A,S,T	-
1045	0.333	c=>t	constant	-
1107	0.333	a=>c	constant	-
1116	0.222	a=>g	P,A	-
1137	0.231	g=>a	constant	-
1140	0.300	g=>a	constant	-
1215	1.000	a->c	constant	-
1266	0.429	t=>c	constant	-
1338	0.333	t=>c	constant	-
1346	0.125	g=>c	A,S,T,C	L
1359	0.286	c->t	P,A,L	-

NODE 48-47

POS	c	NUCΔ	AA	SC
39	0.333	c=>t	constant	-
150	0.231	c=>t	A,P,S	-
159	0.167	a=>g	constant	-
165	0.231	a=>t	A,W	-
549	0.200	g->a	constant	-
603	0.143	g->a	constant	-
741	0.111	t->c	S,C,Y	-
861	0.143	c->t	constant	-
906	0.286	t=>c	D,R	-
1212	0.429	a=>g	constant	-
1269	0.600	t->c	constant	-

1410	0.429	a->g	E,D,A,K,P,Q	-
1420	1.000	a->g	T,V,L,K	-
1421	0.667	c->t	T,V,L,K	-
1425	0.429	a->g	L,V,C	-

NODE 47-46

POS	c	NUCΔ	AA	SC
75	0.125	t=>c	Y,F	-
102	0.429	t=>c	constant	-
117	0.500	a=>g	constant	-
177	0.300	c=>t	R,L,A	-
231	0.286	t->c	constant	-
246	0.333	a=>g	constant	-
321	0.333	t=>c	constant	-
346	0.333	c->a	M,L	L
402	0.500	t->c	constant	-
405	0.222	g->a	constant	-
412	0.200	c=>t	S,L	-
519	0.182	t->c	constant	-
522	0.286	c->t	constant	-
552	0.200	t->c	constant	-
660	0.167	t=>c	constant	-
753	0.188	g->a	L,M,I	L
807	0.250	t=>c	constant	-
834	0.600	t=>c	T,M	-
957	0.400	t->a	R,C	-
963	0.182	t->c	C,S	-
1067	1.000	a=>g	K,R,E	L
1194	0.250	t=>c	S,F,A	-
1206	0.111	t=>c	constant	-
1257	0.500	t=>g	constant	-

NODE 46-45

POS	c	NUCΔ	AA	SC
88	0.143	c->a	E,K,Q,T	L
141	0.333	a=>g	constant	-
162	0.429	a=>g	A,W	-
279	0.167	g=>a	constant	-
284	0.286	a->c	N,D,S,T,E,G	L
741	0.111	c->t	S,C,Y	-
762	0.333	a=>t	A,Q,E,V,H,I	L
957	0.400	a->c	R,C	-
1209	0.286	t=>c	constant	-
1266	0.429	t=>c	constant	-
1362	0.429	a=>g	E,D	-

NODE 66-65

POS	c	NUCΔ	AA	SC
62	0.500	g=>a	R,K,T	L
66	0.167	a->g	L,I	-
88	0.143	c->g	E,K,Q,T	L
144	0.333	g->t	constant	-
153	0.111	g->a	constant	-
162	0.429	a=>g	A,W	-
168	0.273	a->g	constant	-
201	0.250	c=>a	constant	-
207	0.600	t=>g	constant	-
255	0.200	t=>c	Y,C	-
256	0.667	g=>c	G,D,E,N,H	L
271	0.250	c->g	P,A,V,T	L
363	0.250	a=>g	constant	-
378	0.500	a->c	constant	-
408	0.167	a->g	constant	-
450	0.214	t=>c	constant	-
453	0.273	a=>g	constant	-
462	0.429	t=>c	constant	-
486	0.167	a->g	L,S	-
492	0.250	a=>g	constant	-
522	0.286	c->t	constant	-
537	0.429	t->a	constant	-
579	0.375	t->c	constant	-
582	0.167	t=>c	constant	-
618	0.333	a=>g	constant	-
621	0.250	t=>c	constant	-
648	0.200	t=>c	constant	-
666	0.500	a=>c	constant	-
684	0.300	t=>g	A,S	-
690	0.429	t=>c	A,G,T	-
705	0.333	t->c	I,V	-
708	0.200	g=>a	constant	-
762	0.333	a=>c	A,Q,E,V,H,I	L
795	0.250	c->a	V,G	-
807	0.250	t=>c	constant	-
816	0.750	a=>g	constant	-
819	0.250	t=>a	constant	-
882	0.100	c->t	constant	-
912	0.333	a=>g	constant	-
933	0.143	c->t	constant	-
984	0.182	c->t	A,S,T	-
990	0.500	t->a	T,I	-
1005	0.375	t=>g	constant	-
1017	0.333	a->c	constant	-
1020	0.200	a->g	Q,E,D	-
1060	0.333	a->g	Y,F,C,L	-
1107	0.333	a=>c	constant	-
1131	0.333	a->g	constant	-
1206	0.111	t->c	constant	-
1266	0.429	t=>g	constant	-
1278	0.500	t=>g	A,V	-
1330	0.167	g->a	I,V	L
1347	0.200	t->c	A,S,T,C	-
1401	0.250	t=>c	constant	-
1407	0.500	t->c	F,I,L	-
1411	0.600	a->c	T,A,S,E,P	L

NODE 65-64

POS	c	NUCΔ	AA	SC
165	0.231	a=>t	A,W	-
186	0.222	c=>t	constant	-
228	0.125	t=>c	N,S	-
351	0.167	c=>t	constant	-
456	0.222	t->c	constant	-
537	0.429	a->g	constant	-
555	0.100	t->c	constant	-
672	0.300	a->t	constant	-
673	0.111	a=>c	L,I	L
741	0.111	t=>c	S,C,Y	-
753	0.188	g->a	L,M,I	L
879	0.667	t=>c	constant	-
915	0.167	a=>g	K,R	-
982	0.182	g=>t	A,S,T	L
990	0.500	a->c	T,I	-
1011	0.429	a=>g	constant	-
1017	0.333	c->g	constant	-
1047	0.429	g=>a	constant	-
1080	0.333	t=>c	constant	-
1137	0.231	g=>a	constant	-
1167	0.200	t=>c	A,L	-
1194	0.250	t=>c	S,F,A	-
1356	0.143	t->c	constant	-
1411	0.600	c->g	T,A,S,E,P	-
1422	0.429	g=>c	T,V,L,K	-
1425	0.429	a->g	L,V,C	-

NODE 64-63

POS	c	NUCΔ	AA	SC
150	0.231	c=>t	A,P,S	-
153	0.111	a->g	constant	-
309	0.143	t=>c	constant	-
378	0.500	c->g	constant	-
474	0.250	a=>g	constant	-
564	0.214	a->g	A,V	-
612	0.111	a->g	constant	-
696	0.286	g=>a	constant	-
753	0.188	a->c	L,M,I	-
771	0.375	t=>c	constant	-
813	0.231	g->a	constant	-
885	0.286	t=>c	constant	-
927	0.231	g=>a	I,M	L
951	0.222	a->g	constant	-
1060	0.333	g->a	I,Y	-
1299	0.125	a=>g	constant	-
1320	0.143	g=>a	Q,E,A	-
1380	0.200	a=>g	E,A	-

NODE 63-54

POS	c	NUCΔ	AA	SC
84	0.214	g=>c	D,E,Q	L
433	0.250	t->a	T,V,S,I	L
546	0.250	t=>c	constant	-
672	0.300	t->c	constant	-
1020	0.200	g->c	Q,E,D	L

NODE 54-53

POS	c	NUCΔ	AA	SC
543	0.333	t=>c	constant	-
813	0.231	a->g	constant	-
982	0.182	t->g	A,S,T	L
1245	0.200	a->t	constant	-

NODE 53-51

POS	c	NUCΔ	AA	SC
45	0.750	t->c	constant	-
424	0.200	c=>a	V,P,L,T,I	L
425	0.200	c->t	V,P,L,T,I	PNL
433	0.250	a=>g	T,V,S,I	L
434	0.250	c=>t	T,V,S,I	L
672	0.300	c->t	constant	-
753	0.188	c=>g	L,M,I	L
864	0.333	c->t	constant	-
915	0.167	g->a	R,K	-
1408	0.500	g=>a	E,D,A,K,P,Q	L

NODE 51-49

POS	c	NUCΔ	AA	SC
162	0.429	g->a	constant	-
168	0.273	g->a	constant	-
655	0.250	t=>g	L,V,C	L
684	0.300	g=>a	S,A	-
732	0.125	a=>g	constant	-
836	0.222	g=>c	S,I,T	L
1131	0.333	g=>a	constant	-
1167	0.200	c=>t	A,L	-
1345	0.154	a=>t	A,S,T,C	L

NODE 51-50

POS	c	NUCΔ	AA	SC
57	0.333	t=>g	D,E	L
84	0.214	c=>a	D,E,Q	L
284	0.286	a=>g	N,D,S,T,E,G	L
561	0.333	a=>g	constant	-
774	0.500	a=>g	R,K	-
1111	0.286	a=>t	L,M,T	L
1140	0.300	g=>c	constant	-
1318	0.500	g=>c	Q,E,A	L

NODE 53-52

POS	c	NUCΔ	AA	SC
108	0.400	c=>t	I,T	-
290	0.125	a=>t	Y,F	L
297	0.200	t=>c	A,V,C	-
357	0.286	c=>t	constant	-
673	0.111	c=>a	L,I	L
682	0.333	g=>t	S,A	L
771	0.375	c=>t	constant	-
807	0.250	c=>t	constant	-
855	0.400	a=>g	constant	-
1239	0.500	t=>c	constant	-
1410	0.429	a->c	E,D,A,K,P,Q	L

NODE 63-62

POS	c	NUCΔ	AA	SC
138	0.273	c=>t	P,L	-
279	0.167	g=>a	constant	-
435	0.300	c=>t	T,V,S,I	-
456	0.222	c->t	constant	-
732	0.125	a->g	constant	-
762	0.333	c->g	constant	-
861	0.143	c=>t	constant	-
1017	0.333	g->a	constant	-
1032	0.429	t=>c	constant	-
1245	0.200	a->g	constant	-
1251	0.273	a->c	G,A	-
1266	0.429	g->a	constant	-
1270	0.500	t->c	L,S,V	-
1341	0.333	a->g	constant	-
1356	0.143	c->t	constant	-
1422	0.429	c->t	T,V,L,K	-

NODE 62-61

POS	c	NUCΔ	AA	SC
84	0.214	g=>a	D,E,Q	-
165	0.231	t->c	A,W	-
276	0.286	g=>a	constant	-
393	0.231	a=>c	R,P	-
420	0.250	t->c	I,V	-
672	0.300	t->a	constant	-
684	0.300	g=>a	S,A	-
762	0.333	g->t	constant	-
858	0.167	c=>t	constant	-
1005	0.375	g->t	constant	-
1015	0.500	c=>a	constant	-
1111	0.286	a=>c	L,M,T	-
1113	0.500	g=>a	L,M,T	-
1167	0.200	c=>t	A,L	-

NODE 61-59

POS	c	NUCΔ	AA	SC
177	0.300	c->g	R,L,A	-
290	0.125	a=>t	Y,F	L
564	0.214	g=>a	A,V	-
690	0.429	c=>t	A,G,T	-
732	0.125	g->a	constant	-
836	0.222	g->c	S,I,T	L
1278	0.500	g=>a	A,V	-
1401	0.250	c=>t	constant	-

NODE 59-58

POS	c	NUCΔ	AA	SC
153	0.111	g->a	constant	-
177	0.300	g->t	R,L,A	-
228	0.125	c=>t	N,S	-
284	0.286	a=>g	N,D,S,T,E,G	L
312	0.182	t=>c	P,F	-
390	0.667	a=>g	constant	-
450	0.214	c->t	constant	-
528	0.429	a=>t	constant	-
603	0.143	g=>a	constant	-
673	0.111	c=>a	L,I	L
711	0.250	a->g	constant	-
885	0.286	c=>t	constant	-
927	0.231	a=>g	I,M	L
982	0.182	t=>g	A,S,T	L
1137	0.231	a->g	constant	-
1320	0.143	a=>g	Q,E,A	-
1380	0.200	g->a	E,A	-

NODE 58-57

POS	c	NUCΔ	AA	SC
84	0.214	a=>c	D,E,Q	L
147	0.154	a=>g	constant	-
225	0.333	c->t	constant	-
412	0.200	c=>t	S,L	-
498	0.333	t->c	constant	-
543	0.333	t=>c	constant	-
655	0.250	t=>c	L,V,C	-
684	0.300	a=>g	S,A	-
753	0.188	c->g	L,M,I	L
836	0.222	c->g	S,I,T	L
1185	0.200	g=>a	constant	-
1224	0.429	a=>g	constant	-
1251	0.273	c->t	G,A	-

NODE 57-56

POS	c	NUCΔ	AA	SC
60	0.250	c=>t	constant	-
117	0.500	a->c	constant	-
153	0.111	a->g	constant	-
220	0.500	c->g	L,V	L
267	0.375	c->a	P,T	-
378	0.500	g=>a	constant	-
384	0.333	a=>g	constant	-
424	0.200	c->a	V,P,L,T,I	L
486	0.167	g=>a	L,S	-
501	0.333	t=>c	constant	-
510	0.167	g=>a	constant	-
537	0.429	g=>a	constant	-
552	0.200	c=>t	constant	-
588	0.400	a=>g	constant	-
621	0.250	c=>t	constant	-
813	0.231	a=>c	constant	-
864	0.333	c=>t	constant	-
963	0.182	t->c	C,S	-
1029	0.250	g=>a	constant	-
1058	0.500	t=>a	Y,F,C,L	L
1071	0.167	c=>t	constant	-
1077	0.200	t=>c	constant	-
1137	0.231	g->a	constant	-
1176	0.250	g=>a	E,D	-
1203	0.200	g=>a	constant	-
1209	0.286	t=>c	constant	-
1245	0.200	g->c	constant	-
1345	0.154	a=>g	A,S,T,C	L
1346	0.125	g=>c	A,S,T,C	L
1347	0.200	c=>t	A,S,T,C	-
1362	0.429	a=>g	E,D	-
1408	0.500	g=>c	E,D,A,K,P,Q	L
1409	0.250	a=>c	E,D,A,K,P,Q	L

NODE 56-55

POS	c	NUCΔ	AA	SC
117	0.500	c->g	constant	-
165	0.231	c->a	A,W	-
168	0.273	g=>a	constant	-
393	0.231	c=>t	R,P	-
763	0.333	g=>a	A,Q,E,V,H,I	L
1335	0.167	c=>t	constant	-

NODE 61-60

POS	c	NUCΔ	AA	SC
225	0.333	c=>t	constant	-
543	0.333	t=>c	constant	-
741	0.111	c=>t	S,C,Y	-
753	0.188	c->a	L,M,I	-
813	0.231	a->g	constant	-
1026	1.000	t=>c	constant	-
1269	0.600	t=>c	constant	-
1345	0.154	a=>t	A,S,T,C	L

NODE 69-68

POS	c	NUCΔ	AA	SC
18	0.333	g->a	- excluded,	
30	0.250	t->c	constant	-
60	0.250	c->t	constant	-
81	0.333	t->a	constant	-
93	0.400	c=>t	T,P,V	-
96	0.333	a=>g	K,L,S	-
99	0.600	t->c	D,A,E	-
147	0.154	a=>g	constant	-
186	0.222	c->t	constant	-
213	0.250	c=>t	constant	-
282	0.429	a=>c	D,E,K,T	L
306	0.375	t=>a	A,V	-
351	0.167	c=>t	constant	-
366	0.667	t=>g	constant	-
372	0.250	a->t	constant	-
393	0.231	a->c	R,P	-
402	0.500	t=>c	constant	-
424	0.200	c->a	V,P,L,T,I	L
433	0.250	t=>a	T,V,S,I	PNL
434	0.250	c=>t	T,V,S,I	PNL
435	0.300	c->a	T,V,S,I	PNL
441	0.286	t->a	constant	-
444	0.167	c->t	F,C	-
495	0.167	t->c	constant	-
498	0.333	t->a	constant	-
528	0.429	a=>t	constant	-
538	0.400	t=>c	L,I	L
603	0.143	g->a	constant	-
651	0.500	t->c	constant	-
654	0.333	c->t	constant	-
655	0.250	t=>g	L,V,C	L
657	1.000	a=>t	L,V,C	-
666	0.500	a->g	constant	-
684	0.300	t->a	S,A	-
720	0.200	g->a	constant	-
732	0.125	a->g	constant	-
753	0.188	g->a	L,M,I	L
768	0.167	t->c	C,F	-

771	0.375	t=>a	constant	-
792	0.500	c->t	I,S	-
804	0.250	c->t	constant	-
834	0.600	t->c	T,M	-
836	0.222	g=>c	S,I,T	L
858	0.167	c=>t	constant	-
865	0.167	c->t	L,P	-
942	0.200	a=>g	L,S	-
957	0.400	t->a	R,C	-
981	0.143	c->t	constant	-
1038	0.500	t=>g	constant	-
1128	0.222	c->t	constant	-
1192	1.000	t=>g	S,F,A	L
1218	0.200	t->c	constant	-
1227	0.286	c->t	H,Q	-
1245	0.200	a->c	constant	-
1251	0.273	a->c	G,A	-
1323	0.333	t->g	constant	-
1332	0.500	t=>a	I,V	-
1410	0.429	a->g	E,D,A,K,P,Q	-
1414	0.333	a->g	T,A,S,E,P	L

NODE 68-67

POS	*c*	NUCΔ	AA	SC
33	0.500	t=>a	V,S,F,D,A	L
40	1.000	a=>c	K,Q	L
81	0.333	a->g	constant	-
150	0.231	c=>t	A,P,S	-
207	0.600	t=>g	constant	-
255	0.200	t=>c	Y,C	-
259	1.000	a=>c	I,L	L
261	0.167	c->t	I,L	-
339	0.500	t=>g	constant	-
369	0.250	t=>c	constant	-
387	0.333	c=>t	constant	-
393	0.231	c->g	R,P	-
397	0.333	c->t	L,S,I	-
427	0.333	g=>t	A,S	-
450	0.214	t->a	constant	-
459	0.250	t=>c	constant	-
513	0.750	a->g	constant	-
543	0.333	t->a	constant	-
564	0.214	a->c	A,V	-
618	0.333	a->c	constant	-
633	0.200	t=>c	constant	-
639	0.333	t=>c	constant	-
708	0.200	g=>a	constant	-
717	0.200	c=>t	constant	-
744	0.250	a=>g	constant	-
753	0.188	a->c	L,M,I	-
822	0.143	c->t	constant	-
930	0.111	c=>t	constant	-
969	0.429	a=>g	constant	-
984	0.182	c->t	A,S,T	-
1050	0.400	t=>c	constant	-
1122	0.400	t=>c	constant	-
1176	0.250	g->a	E,D	-
1179	0.400	c=>t	constant	-
1251	0.273	c->t	G,A	-
1287	0.200	a=>g	E,Q,K	-
1302	0.286	a=>g	constant	-
1320	0.143	g=>a	Q,E,A	-
1338	0.333	t=>c	constant	-
1374	1.000	t->g	constant	-
1383	0.500	a->c	V,I	-
1411	0.600	a->t	T,A,S,E,P	L
1413	1.000	a->g	T,A,S,E,P	-
1422	0.429	g->a	T,V,L,K	-
1425	0.429	a->g	L,V,C,I	-

NODE 76-75

POS	*c*	NUCΔ	AA	SC
28	0.500	a->g	- excluded,	
61	1.000	a=>c	R,K,T	-
66	0.167	a=>g	L,I	-
90	0.250	g=>a	E,K,Q,T	-
189	0.375	t->c	constant	-
195	0.750	a->c	constant	-
207	0.600	t=>a	constant	-
213	0.250	c->a	constant	-
345	0.333	c->t	constant	-
375	0.333	t=>c	F,S	-
403	0.333	c=>t	constant	-
415	0.250	c=>a	constant	-
445	0.500	c->a	Q,M,I,T,L,W	L
446	0.500	a->t	Q,M,I,T,L,W	NL
459	0.250	t=>c	constant	-
600	0.333	t->c	constant	-
612	0.111	a=>g	constant	-
677	0.167	a=>t	Y,N,F	L
684	0.300	t->c	S,A	-
693	0.167	a->g	constant	-
723	0.250	t->c	constant	-
764	0.400	a=>t	A,Q,E,V,H,I	NL
765	0.429	a=>c	A,Q,E,V,H,I	L
808	0.167	t->c	constant	-
845	0.500	a->t	H,Y,S,F	L
897	0.667	a=>t	A,V	-
906	0.286	c->t	D,R	-
996	0.400	a->g	constant	-
1032	0.429	t->a	constant	-
1080	0.333	t->c	constant	-
1122	0.400	t=>a	constant	-
1123	0.250	t->c	L,S,F,I,M	NL
1140	0.300	a->t	constant	-
1236	0.333	a->t	constant	-
1275	0.333	a->g	constant	-

NODE 75-74

POS	*c*	NUCΔ	AA	SC
31	0.500	g->t	V,S,F,D,A	PNL
96	0.333	a=>g	K,L,S	-
108	0.400	t=>c	I,T	-
109	0.333	t=>c	constant	-
159	0.167	a->g	constant	-
195	0.750	c->g	constant	-
228	0.125	t=>c	N,S	-
261	0.167	t=>c	I,L	-
276	0.286	a->c	constant	-
315	0.167	a->g	constant	-
339	0.500	t->c	constant	-
444	0.167	t=>c	F,C	-
453	0.273	a->g	constant	-
522	0.286	t=>c	constant	-
618	0.333	a->c	constant	-
738	0.667	a->g	T,N	-
780	0.143	a->g	constant	-
786	0.500	t->a	V,M,A	-
789	0.429	t=>c	constant	-
795	0.250	t->c	V,G	-
801	0.500	t=>c	constant	-
819	0.250	t=>a	constant	-
914	0.143	a->g	K,R	L
927	0.231	t->g	I,M	L
954	0.286	a=>g	S,L	-
976	0.250	a->g	constant	-
1008	0.500	a->g	constant	-
1032	0.429	a->c	constant	-
1035	0.250	t->c	constant	-
1086	1.000	t=>c	I,V,L	-
1101	0.250	c->t	constant	-
1116	0.222	t=>g	P,A	-
1182	0.250	t->c	constant	-
1287	0.200	a->g	Q,E,K	-
1350	0.250	a=>g	constant	-
1365	0.333	a->g	constant	-
1395	1.000	c=>t	constant	-
1401	0.250	t=>c	constant	-

NODE 74-73

POS	*c*	NUCΔ	AA	SC
69	0.500	t=>c	N,T	-
88	0.143	g->a	E,K,Q,T	L
120	0.167	c->t	constant	-
153	0.111	a=>g	constant	-
201	0.250	t->c	constant	-
213	0.250	a->c	constant	-
225	0.333	t=>c	constant	-
243	0.200	a->g	constant	-
276	0.286	c->g	constant	-
306	0.375	t->a	A,V	-
327	0.167	a->g	constant	-
345	0.333	t->c	constant	-
351	0.167	t=>c	constant	-
387	0.333	t=>a	constant	-
390	0.667	a=>g	constant	-
391	0.333	c=>a	R,P	-
397	0.333	t->c	L,S,I	L
405	0.222	a=>g	constant	-
423	0.500	t->c	constant	-
438	0.333	a=>g	K,Q	-
447	0.167	a=>g	Q,M,I,T,L,W	L
528	0.429	a->c	constant	-
678	1.000	t=>c	Y,N,F	-
693	0.167	g->a	constant	-
708	0.200	a=>g	constant	-
738	0.667	g->t	T,N	-
772	0.500	a=>c	R,K	-
808	0.167	c->t	constant	-
846	0.333	t->c	H,Y,S,F	-
942	0.200	a=>g	L,S	-
1023	0.231	a->c	V,I,L,M	-
1027	0.167	t->c	constant	-
1056	0.167	c->t	constant	-
1065	0.333	a=>g	constant	-
1077	0.200	t=>c	constant	-
1083	0.600	t->c	constant	-
1140	0.300	t->a	constant	-
1179	0.400	t=>c	constant	-
1185	0.200	a->g	constant	-
1221	0.200	a=>g	L,S	-
1254	0.429	a->g	constant	-
1272	0.667	a=>g	L,S,V	-
1275	0.333	g->a	constant	-
1359	0.286	t=>c	P,A,L	-
1398	0.250	a->g	R,K,I	-

NODE 73-72

POS	*c*	NUCΔ	AA	SC
31	0.500	t->g	V,S,F,D,A	PNL
81	0.333	t=>c	constant	-
117	0.500	a->c	constant	-
132	0.286	t=>c	constant	-
144	0.333	a=>g	constant	-
148	0.333	c=>g	A,P,S	L

```
168  0.273 a=>g constant      -
204  0.375 t=>c constant      -
258  0.333 t=>a G,D,E,N,H     L
291  0.333 t=>c Y,F           -
306  0.375 a->g A,V           -
313  0.250 t=>c constant      -
315  0.167 g->a constant      -
321  0.333 a->c constant      -
402  0.500 t=>g constant      -
465  0.333 c=>t constant      -
471  0.500 t=>c V,A           -
477  0.143 a=>g constant      -
504  0.167 t=>c constant      -
534  0.200 a=>g constant      -
537  0.429 t=>a constant      -
546  0.250 t=>c constant      -
552  0.200 c=>t constant      -
577  1.000 c=>t constant      -
579  0.375 t=>a constant      -
588  0.400 a=>g constant      -
591  0.667 t=>c constant      -
597  0.167 t=>c constant      -
603  0.143 a=>g constant      -
612  0.111 g=>a constant      -
618  0.333 c->a constant      -
663  0.500 a=>g V,C           -
696  0.286 a=>g constant      -
702  0.200 a=>g constant      -
729  0.250 t=>c constant      -
732  0.125 a=>g constant      -
744  0.250 a=>g constant      -
753  0.188 g=>a L,M,I         L
765  0.429 c->a C,F           -
785  0.200 c=>t V,M,A         L
804  0.250 c=>t constant      -
837  0.300 t=>c S,I,T         -
840  0.167 g=>a L,S           -
849  0.250 t=>c constant      -
852  0.286 t=>c constant      -
870  0.600 t=>c constant      -
876  0.143 t=>c constant      -
914  0.143 g->a K,R           L
945  0.750 t=>c constant      -
969  0.429 a=>g constant      -
976  0.250 g->a constant      -
1021 0.333 g=>a V,I,L,M       L
1042 0.167 t=>c L,S           -
1053 0.200 t=>c constant      -
1095 0.500 t=>c constant      -
1194 0.250 t=>c S,F,A         -
1198 0.167 t=>c L,S           -
1200 0.600 a->c L,S           -
1230 0.667 t->a constant      -
1345 0.154 a->g A,S,T,C       L
1346 0.125 g=>c A,S,T,C       L
1362 0.429 a->c E,D           L
1365 0.333 g->a constant      -
1380 0.200 a=>g E,A           -
1392 0.143 a=>g constant      -
```

APPENDIX II (pp. 562–567; corrections in proof, p. 566). Inferred amino acid changes on the internal branches of a string-based cladogram (one of 165 equally most-parsimonious), including summary statistics of the string search and the resultant matrix of apomorphic recognitions.

Similar to Appendix I, the following table and accompanying reference cladogram contain information about the functional impact of specific string changes (as reflected by alterations in amino acid identity). Interpretation is as in Appendix I with the following exceptions: (i) relative branch length (changes per given branch divided by total steps) is given, (ii) "CHAR" indicates the string character number from the matrix at the end of this appendix, (iii) "POS." still refers to nucleotide position, but, here, to the starting (3′) position of a string recognition, (iv) "STR., SEQ." indicates first the number of simulated nucleotides (i.e., string length) followed by the string itself (divided to show the codon positions of its component nucleotides), and (v) "AA-seq." shows each alternative amino acid *sequence* identified by a particular string recognition. Under the latter category, internal stop codons are indicated by *1, *2, or *3 (for TAA, TAG, and TGA, respectively), and missing nucleotide data have sometimes necessitated the indication (by "?") of missing amino acids. Again, Dayhoff et al. (1978) PAM-250 log-odds calculations were determined nondirectionally for each combination of amino acid sequences.

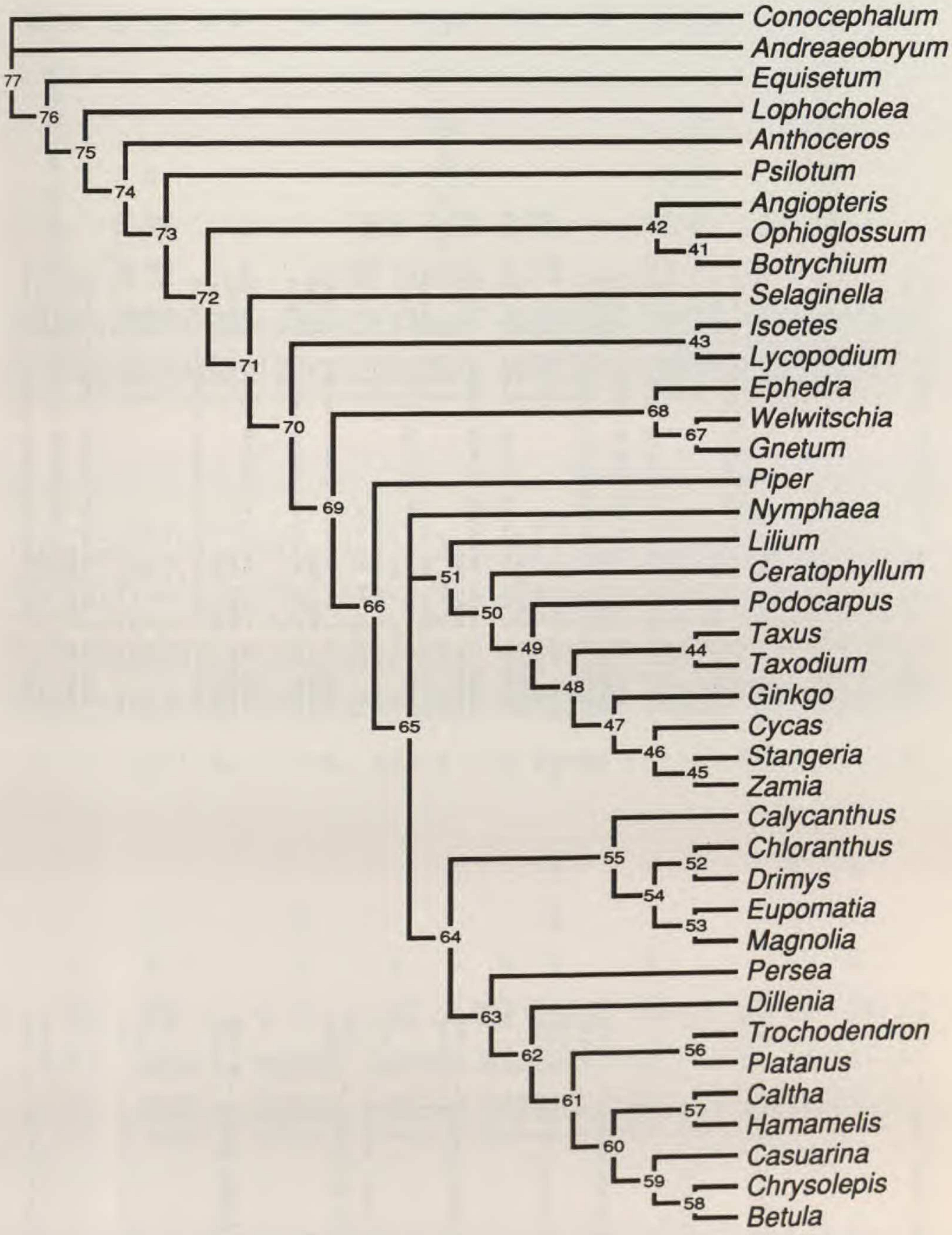
Summary statistics from the string search (involving 1000 randomly generated strings ranging in length from 6 to 21 base pairs) are provided below.

String length	Total recognitions	Total apomorphies	Total similarities	Total singletons	Total positional recognitions	Mean recognitions per string
6	758	129	77	52	47	2.745
7	204	43	20	23	31	1.387
8	107	14	10	4	13	1.077
9	5	2	1	1	2	1.000
10	4	2	1	1	2	1.000
12	21	1	1	0	1	1.000
14	5	1	1	0	1	1.000
15	8	1	1	0	1	1.000
Σ	1112	193	112	81	98	

The 1000 strings evaluated contained the following proportions of "nucleotides," which verify their random generation:

$$\Sigma\ A = 3375$$
$$\Sigma\ C = 3309$$
$$\Sigma\ G = 3349$$
$$\Sigma\ T = 3297$$

The matrix of 193 string recognitions (including 112 potentially informative similarities) is also presented. Headers are provided to give additional information for each character. The number of nucleotides per string character is given, followed by the number of recognitions (hits) per string, the start position of the string (in terms of *rbc*L nucleotides), and the character number (for reference to the table of changes). Immediately following the start position information may appear the designation "ab"; this indicates that separate string recognitions had the same start position, and so showed partial overlap (such partial correlation has been ignored in our present analyses; see text for further details). The matrix is presented in two blocks, corresponding to two rounds of string evaluation (500 strings in each, for a total of 1000). In each case, string recognitions occurring in the 3′ primer region are shown in brackets, but were ignored during parsimony analysis.


Conocephalum
Andreaeobryum
Equisetum
Lophocholea
Anthoceros
Psilotum
Angiopteris
Ophioglossum
Botrychium
Selaginella
Isoetes
Lycopodium
Ephedra
Welwitschia
Gnetum
Piper
Nymphaea
Lilium
Ceratophyllum
Podocarpus
Taxus
Taxodium
Ginkgo
Cycas
Stangeria
Zamia
Calycanthus
Chloranthus
Drimys
Eupomatia
Magnolia
Persea
Dillenia
Trochodendron
Platanus
Caltha
Hamamelis
Casuarina
Chrysolepis
Betula
77
76
75
74
73
72
71
70
69
66
65
64
63
62
61
60
59
58
57
56
55
54
53
52
51
50
49
48
47
46
45
44
43
42
41
68
67

```
NODE 77 - 76, relative branch length = 0.0138
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
032 313b 7, tta gat t            0.500  LDL                                -
091 1254 6, t gct aa             1.000  VAN, AAN                           L
100 1344 6, t gct aa             0.200  AAK, ASK, ACK, ATK, RTK          PNL
136  465 6, t caa gt             0.250  IQV                                -
142  607 6, gat gaa              0.125  DE                                 -
172  980 7, ac gct gg            0.100  HAG, HSG, HTG                      L
173 1017 6, t caa gt             0.500  RQV, RDV, REV, REI, RQI, RDL       L

NODE 76 - 75, relative branch length = 0.0138
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
033  326 6, aa gaa g             0.125  EEG                                -
043  487 6, aac aaa              0.143  NK                                 -
113  111 6, a gca gc             0.200  LAA                                -
121  235 6, cgt tac              0.333  RY                                 -
137  728 6, ct gca g             0.167  TAG, TSG                           L
152  750 6, g atg aa             0.100  MMK, MLK, MIK                      L
183 1231 6, tgg gga              0.200  WG                                 -

NODE 75 - 74, relative branch length = 0.0138
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
035  345 6, c atg tt             0.200  NMF, NLF                           L
100 1344 6, t gct aa             0.200  AAK, ASK, ACK, ATK, RTK          PNL
116  162 6, a gca gc             0.250  GAA, GWA                          NL
150  724 6, gct act              0.125  AT                                 -
158  830 8, at act agt           0.167  NTS, NMI, NTT                    PNL
166 1259 7, at cga gt            0.143  NRV, N*3V                          -

NODE 74 - 73, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
018  152 6, aa gaa g             0.143  EEA                                -
031  313 6, tta gat              0.500  LD                                 -
043  487 6, aac aaa              0.143  NK                                 -
092 1259 7, at cga gt            0.167  NRV                                -

NODE 73 - 72, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
034  333 8, t tct gtt a          0.167  GSVT                               -
085 1147 6, cat gtt              0.143  HV                                 -
138  500 7, gt cct tt            0.250  RPL                                -
183 1231 6, tgg gga              0.200  WG                                 -

NODE 72 - 42, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
033  326 6, aa gaa g             0.125  EEG                                -
115  155 7, aa gca gg            0.167  EAG                                -
137  728 6, ct gca g             0.167  TAG, TSG                           L
187 1284 6, a cag gc             0.250  VQA, VEA, VKA                      L
```

```
NODE 42 - 41, relative branch length = 0.0178
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
044  543 6, t gct aa             0.143  SAK                                -
052  728 6, ct gca g             0.111  TAG, TSG                           L
068  939 7, a ttg gcc            1.000  VLA, VSA                          NL
077 1093 8, acc caa ga           0.250  TQD, PQD                           L
100 1344 6, t gct aa             0.200  AAK, ASK, ACK, ATK, RTK          PNL
113  111 6, a gca gc             0.200  LAA                                -
136  465 6, t caa gt             0.250  IQV                                -
142  607 6, gat gaa              0.125  DE                                 -
152  750 6, g atg aa             0.100  MMK, MLK, MIK                      L

NODE 72 - 71, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
026  266 6, ct gtt g             0.167  PVA, PVP, PVV, TVT, SVV          PNL
053  755 7, aa aga gc            0.200  KRA                                -
114  126 6, g act cc             0.500  MTP, VTP, LTP, VSP                 L
193 1394 6, tc aag t             0.500  IKF, IRF, IIF                    PNL

NODE 71 - 70, relative branch length = 0.0138
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
034  333 8, t tct gtt a          0.167  GSVT                               -
044  543 6, t gct aa             0.143  SAK                                -
052  728 6, ct gca g             0.111  TAG, TSG                           L
092 1259 7, at cga gt            0.167  NRV                                -
142  607 6, gat gaa              0.125  DE                                 -
152  750 6, g atg aa             0.100  MMK, MLK, MIK                      L

NODE 70 - 43, relative branch length = 0.0099
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
006   88 6, aag acc              0.500  ET , EP, KV, DT, QT, TP          PNL
007   90 6, g acc aa             0.200  ETK, EPK, KVS, KTK, DTK, QTK, ETL,
                                        TPK, PNL
114  126 6, g act cc             0.500  MTP, VTP, LTP, VSP                 L
138  500 7, gt cct tt            0.250  RPL                                -
158  830 8, at act agt           0.167  NTS, NMI, NTT                    PNL

NODE 70 - 69, relative branch length = 0.0039
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
033  326 6, aa gaa g             0.125  EEG                                -
183 1231 6, tgg gga              0.200  WG                                 -

NODE 69 - 66, relative branch length = 0.0138
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
013  141 6, a gtt cc             0.333  GVP                                -
036  388 6, cta cga              0.200  LR, LP                             L
043  487 6, aac aaa              0.143  NK                                 -
056  783 6, a gtt cc             0.250  GVP, GMP, GAP                    PNL
124  273 6, t ggg ga             0.167  AGE, PGE, VGE, TGE               PNL
177 1182 6, t ggg ga             0.333  FGD                                -
193 1394 6, tc aag t             0.500  IKF, IRF, IIF                    PNL

NODE 66 - 65, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                   c    AA-seq.                          SC
010  123 12, a gta act cct ca    0.200  RVTPQ, RMTPQ, RLTPQ, RVSPQ         L
076 1067 6, aa gac c             0.200  KFR, EDR                          NL
132  395 6, ct cta c             0.143  ALR, ASR, T??                     NL
150  724 6, gct act              0.125  AT                                 -
```

```
 NODE 65 - 51, relative branch length = 0.0059
CHARPOS. STR.,SEQ.                   c    AA-seq.                              SC
085 1147 6, cat gtt              0.143  HV                                    -
152  750 6, g atg aa             0.100  MMK, MLK, MIK                         L
166 1259 7, at cga gt            0.143  NRV, N*V                              -

 NODE 51 - 50, relative branch length = 0.0059
CHARPOS. STR.,SEQ.                   c    AA-seq.                              SC
124  273 6, t ggg ga             0.167  AGE, PGE, VGE, TGE                   PNL
130  386 6, ct cta c             0.167  ALR, ALP                              L
142  607 6, gat gaa              0.125  DE                                    -

 NODE 50 - 49, relative branch length = 0.0118
CHARPOS. STR.,SEQ.                   c    AA-seq.                              SC
013  141 6, a gtt cc             0.333  GVP                                   -
035  345 6, c atg tt             0.200  NMF, NLF                              L
047  635 6, tg cgt t             0.333  MRW                                   -
051  684 6, t cag gc             0.250  AQA, SQA, AQT, SQG                    L
053  755 7, aa aga gc            0.200  KRA                                   -
143  639 15, c tgg aga gat cgt tt0.500  RWRDRF                                -

 NODE 49 - 48, relative branch length = 0.0158
CHARPOS. STR.,SEQ.                   c    AA-seq.                              SC
050  663 6, t gca ga             0.500  CAE, VAE, CAE                        NL
052  728 6, ct gca g             0.111  TAG, TSG                              L
085 1147 6, cat gtt              0.143  HV                                    -
115  155 7, aa gca gg            0.167  EAG                                   -
119  199 6, acc act              0.200  TT                                    -
130  386 6, ct cta c             0.167  ALR, ALP                              L
137  728 6, ct gca g             0.167  TAG, TSG                              L
166 1259 7, at cga gt            0.143  NRV, N*V                              -

 NODE 48 - 44, relative branch length = 0.0138
CHARPOS. STR.,SEQ.                   c    AA-seq.                              SC
031  313 6, tta gat              0.500  LD                                    -
038  412 6, cta cga              0.333  LR, SR                               NL
116  162 6, a gca gc             0.250  GAA, GWA                             NL
138  500 7, gt cct tt            0.250  RPL                                   -
142  607 6, gat gaa              0.125  DE                                    -
172  980 7, ac gct gg            0.100  HAG, HSG, HTG                         L

 NODE 48 - 47, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                   c    AA-seq.                              SC
017  164 6, ct gca g             0.167  AAV, WAV                             NL
053  755 7, aa aga gc            0.200  KRA                                   -
054  766 7, ttt gcc a            0.250  FAR, CAR, CAK                       PNL
124  273 6, t ggg ga             0.167  AGE, PGE, VGE, TGE                  PNL
```

```
 NODE 47 - 46, relative branch length = 0.0158
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
002   54 14, a gat tac aga tta 0.200KDYRL, RDYRL, KDYKL, KDYTI, KEYKL PNL
024  227 6, gt ctc g             0.250  SLD, NLD                                  L
035  345 6, c atg tt             0.200  NMF, NLF                                  L
043  487 6, aac aaa              0.143  NK                                        -
049  655 6, tgc ttc              1.000  CF, LF, VF                               NL
076 1067 6, aa gac c             0.200  KFR, EDR                                 NL
092 1259 7, at cga gt            0.167  NRV                                       -
152  750 6, g atg aa             0.100  MMK, MLK, MIK                             L

 NODE 46 - 45, relative branch length = 0.0059
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
012  140 7, gg gtg cc            0.333  GVP                                       -
132  395 6, ct cta c             0.143  ALR, ASR, T??                            NL
1821207b 6, ggc ggg              0.500  GG                                        -

 NODE 65 - 64, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
052  728 6, ct gca g             0.111  TAG, TSG                                  L
142  607 6, gat gaa              0.125  DE                                        -
168  950 6, cg tta c             0.250  ALR, ASC                                 NL
172  980 7, ac gct gg            0.100  HAG, HSG, HTG                             L

 NODE 64 - 55, relative branch length = 0.0059
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
061  856 6, gac aac              0.200  DN                                        -
106 1418 6, at acc t             0.500  DTL, DVL, DTV, ILC, DKL                  NL
191 1355 8, gc cct gaa           0.500  SPE, SPD, SAE, SLE                        L

 NODE 55 - 54, relative branch length = 0.0039
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
054  766 7, ttt gcc a            0.250  FAR, CAR, CAK                            NL
152  750 6, g atg aa             0.100  MMK, MLK, MIK                             L

 NODE 54 - 52, relative branch length = 0.0020
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
021  198 6, g aca ac             0.250  WTT                                       -

 NODE 54 - 53, relative branch length = 0.0059
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
073 1135 6, tca ggc              0.500  SG                                        -
079 1110 7, t ttg cca            1.000  SLP, STP, SLA, SMP                       NL
176 1138 6, ggc ggt              0.500  GG                                        -

 NODE 64 - 63, relative branch length = 0.0020
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
192 1369 7, gct gct t            0.167  AAC                                       -

 NODE 63 - 62, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                   c    AA-seq.                                  SC
003   74 9, at acg cct g         0.200  YTPE, FTPD, YTPQ, YTPD                    L
054  766 7, ttt gcc a            0.250  FAR, CAR, CAK                            NL
092 1259 7, at cga gt            0.167  NRV                                       -
124  273 6, t ggg ga             0.167  AGE, PGE, VGE, TGE                      PNL
```

```
NODE 62 - 61, relative branch length = 0.0039
CHARPOS. STR.,SEQ.                  c     AA-seq.                        SC
098 1338 6, t gag gc              0.333  REA, EEA                        NL
175 1109 6, ct cta c              0.333  SLP, STP, SLA, SMP              NL

NODE 61 - 56, relative branch length = 0.0039
CHARPOS. STR.,SEQ.                  c     AA-seq.                        SC
015  146 8, ca cct gag            0.100  PPE, PAE, PSE                  PNL
145  684 6, a cag gc              0.333  AQA, SQA, SQG, AQT               L

NODE 60 - 57, relative branch length = 0.0039
CHARPOS. STR.,SEQ.                  c     AA-seq.                        SC
018  152 6, aa gaa g              0.143  EEA                              -
090 1245 7, g ggt gcc             0.500  PGA, PGG, PVA, PGR, PRA         NL

NODE 60 - 59, relative branch length = 0.0099
CHARPOS. STR.,SEQ.                  c     AA-seq.                        SC
076 1067 6, aa gac c              0.200  KFR, EDR                        NL
100 1344 6, t GCT aa              0.200  AAK, ASK, ACK, ATK, RTK        PNL
137  728 6, ct gca g              0.167  TAG, TSG                         L
152  750 6, g atg aa              0.100  MMK, MLK, MIK                    L
177 1182 6, t ggg ga              0.333  FGD                              -
```

```
NODE 59 - 58, relative branch length = 0.0039
CHARPOS. STR.,SEQ.                  c     AA-seq.                        SC
139  501 6, c ccc ct              1.000  RPL                              -
146  686 9, ag gct gaa a          0.333  QAET, QGET, QTET                 L

NODE 69 - 68, relative branch length = 0.0079
CHARPOS. STR.,SEQ.                  c     AA-seq.                        SC
027  267 6, t gtt cc              0.250  PVP, PLP, PVV, TVT, SVV, PVA   PNL
029  543 6, t gct aa              1.000  SAK                              -
137  728 6, ct gca g              0.167  TAG, TSG                         L
166 1259 7, at cga gt             0.143  NRV, N*V                         -

NODE 68 - 67, relative branch length = 0.0197
CHARPOS. STR.,SEQ.                  c     AA-seq.                        SC
025  252 6, c tac ga              0.250  CYD, CYG, CYE, CYN, CYH        PNL
034  333 8, t tct gtt a           0.167  GSVT                             -
047  635 6, tg cgt t              0.333  MRW                              -
076 1067 6, aa gac c              0.200  KFR, EDR                        NL
092 1259 7, at cga gt             0.167  NRV                              -
116  162 6, a gca gc              0.250  GAA, GWA                        NL
130  386 6, ct cta c              0.167  ALR, ALP                         L
150  724 6, gct act               0.125  AT                               -
170  966 6, t ggg ga              0.333  GGD                              -
192 1369 7, gct gct t             0.167  AAC                              C
```

Corrections in proof: P. 564, under"NODE 70–43," fourth line from bottom, delete comma after "TPK" and move "PNL" to right hand column; p. 566, under "NODE 68–67," bottom line, right hand column, should read "—."

```
#nucl./string  76 8197666761 6767886666 6776666616 6768666666 6666676666 6677667866 6667667767 6678668677 6676668767 6776676676 6667666    766 787 6666766667 6686676666 6667666766 6616696766 6666766867 7676666666 6767666176 6668676866 876
               -- -4-------2 ---------- --------0- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -------    --- --- ---------- ---------- ---------- --5------- ---------- ---------- -------0-- ---------- ---
Σ#hits/string  78 3514228272 3743113438 1225216438 9443422241 1212213138 9312242122 1391122232 4131425322 7123412217 2223312822 2294871    311 321 1311311391 4837215473 4312511152 1382141183 6311281192 2328128467 1111317118 1671121313 621
               -- 3---56---1 --2-1-505- 7-0-58-0-- --0145-528 -6745-0-0- -52042---4 70--7---0- 48--06--00 -8-047--8- 40-00----4 -------    387 3-- -2913908-- 9--10--918 98---482-0 04--7-7--2 -4099-70-- -2--70-9-1 -54-801--- 0-1-9-7-92 -94

String start   00 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0011111111 1111111111 1111111111 1111111    000 000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0011111111 1111111111 111
position       00 0000000011 1111111111 1222222222 3333333444 4445556666 6777777888 8899999999 9910000011 1111122222 2222222333 3333444    000 000 0011111112 2222233333 3344444555 5666666677 7777778888 8888999999 9900111111 2222222333 333
               01 4578889902 3444466577 9112566788 1123489112 2384673556 8256688344 5600011356 7834669912 2233422244 5566689344 7999011    222 467 8912566891 4457800368 9911268004 6038889922 2555661333 5678013566 7817038889 0033558244 569
               48 9441480643 6015604234 8257267803 3363582272 5673735253 4856835319 6925703934 2950473502 3711735545 4936663814 8044789    335 146 7716525399 5943049626 5529156017 1790464524 9099071022 8404537046 8079982495 7715144512 594
               -- ---------- ---------- ---------- ab-------- ---------- ---------- ---------- ---------- --ab--ab-- ---ab----- --ab---    ab- --- ---------- ---------- ab-------- ---------- --ab----ab ---------- ---------- ab-------- ---
Character #    XX 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000001 1111111    XXX 111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 111
               XX 0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990 0000000    XXX 001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 999
               XX 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567    XXX 890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 123

Conocephalum  [??] ??00000000 0000000100 0000001000 1110000000 0011001000 0110000001 0000000000 0100000000 0000100000 1000000000 0??????   [???] ??0 0011110000 0000000000 1000011101 0100000001 0001010100 0000010000 0110000000 1000000010 01?
Lophocolea    [??] ?011000000 1001000100 0000010000 1001000000 000100???? 0110000000 0010000000 ?00?000000 0000100010 0000000001 0??????   [???] ??0 0001110001 1000000000 0000000101 ???0000001 0100111100 0000010000 0000000000 0010010000 01?
Anthoceros    [??] ???0001000 0100000100 0000110000 1101100000 0000?????? ?100000000 0000000000 0?0?000000 0000100000 0000000000 0??????   [???] ?00 0001100000 1000010000 000001010? 0??00?0000 0100000000 0001000000 00?0000000 0110000100 01?
Andreaeobryum [??] ??00011000 0000001011 1000110000 111100?00 0011001000 0110000000 0000000000 0010000000 0000100000 1000000100 ???????   [???] ??0 0011100000 0000000000 0000011100 0100000001 0000000100 0000010000 0110010000 0010000000 0??
Ophioglossum  [10] 1000000000 0000000000 0001010100 0010100000 0010001000 0010000000 0000000100 0000001000 0000000100 0100000001 0011???   [100] 000 0011000000 1001000000 0100011000 0100000000 0000000000 0000010001 0000000000 0000001000 010

Psilotum      [10] 0000000000 0000101000 0000010000 0001100000 0111001000 0110000010 0000000000 0000000000 0101100000 0100100000 0000???   [000] 000 1011100010 1001000000 0101100100 0000000000 0100000100 0000000000 0001001000 0010000000 010
Isoetes       [?0] 1100011000 0000100000 0000000000 0001100000 1010001000 0000000000 0000001000 0000011000 0000001000 0001000000 0000???   [000] 010 0001100010 1000000000 0100000100 0100000000 0110000100 0000000001 0000000000 0000000000 011
Lycopodium    [10] 0000011100 0000000000 0001010000 0000000000 0010001000 0000100000 0100000000 0000000000 0000000000 0000000000 0000???   [010] 000 0001000000 0100100100 0000000100 0100001000 0000010100 0000000000 0000000000 0000001000 001
Angiopteris   [10] 1001000000 0000001100 0100010000 0010000000 0001001000 0110001000 0000110000 1000000000 0000000000 0101000000 0000???   [001] 000 0001000000 1000000000 0000001000 0000001001 0100000001 0000000000 0000000000 0000001000 010
Equisetum     [11] 1000000000 0000000100 0100010000 1011001000 0011001000 0110000000 0000000000 0100000000 0000000000 0000000001 0000???   [100] 000 0011110000 0000000001 0000001100 0000000001 0000000100 0000010000 0000000000 0000000100 010

Selaginella   [00] 1100100000 0000000000 0000000001 0000100000 0111101000 1100000000 0000000000 0000000100 0010000000 0100011000 0000???   [000] 001 0000100000 1000001100 0011000000 0000000101 0100000000 0100100010 1110000000 0000000000 001
Botrychium    [10] 1000000000 0000000000 0000001000 0000100000 0010001000 0010000111 0000000100 0000001000 0000101000 0100000001 0000???   [100] 000 0011000000 1000000000 0000111001 0100000000 0000000000 0000000010 0000000000 0000001000 010
Taxus         [??] 1?00001000 0000000000 0000000000 10?0010100 0010000001 1110010000 1000000000 0000010000 0000000000 0000000000 ???????   [000] 000 0000010010 1000000000 0100001101 0110000001 0100000000 0000000000 0100001000 0010000000 ???
Taxodium      [10] 1000000001 0010000100 1000000000 1011000100 0000000001 1110010000 0000000000 0000010000 0000000000 0000000000 ???????   [000] 000 0000110000 1010000000 0100011100 0110000001 0100000000 0000000000 0100001000 0010010001 0??
Podocarpus    [00] 1100001001 0100000000 0000001000 0011000000 0000000000 1010010000 1000000000 0000010000 1000100000 0000000000 0000000   [000] 000 0000100000 1000000001 0100000000 0010010000 0100000000 0000010000 0000001000 0010000000 01?

Ginkgo        [??] 1000000001 0000101000 0000000000 0001010000 0000000101 0101010000 0000000000 0000010000 0000000000 00?0000000 ???????   [000] 000 0000000010 1001000000 0100000001 0010000001 0100000000 0000001000 0000001000 00??000000 ??0
Cycas         [??] 1100000001 0000001000 0001000000 0011110010 0010000011 1101000000 0000000000 0000001000 0000000000 0100000000 0000100   [000] 000 0010000010 1001100000 0100001000 0010000001 0000000000 0000000000 0000001000 0010000000 010
Stangeria     [??] 1100000000 0100101000 0001000000 0011100000 0010000011 0101010000 0000000000 0000000000 0000000000 0100000000 0000000   [?0?] 000 0000000010 1001000000 0000001000 0010000001 0000000000 0000000000 0000001100 0110000000 010
Zamia         [??] 1000000001 0100001000 0000000000 0011110000 0010010011 1001010000 0000000000 1000000000 0000000000 0100000000 0000000   [?0?] 000 0000000010 1001000000 0000001000 0010000000 0000000000 0000000000 0000001000 0110100000 010
Ephedra       [??] 1000000001 0001000000 0000001010 0011100000 0011001000 0010000000 0000000000 0000000000 0101000000 0000000000 0000001   [?00] 000 0000101000 1000000010 0100001100 0100000000 0000000010 0000010000 0000000000 0010000000 011

Welwitschia   [00] 1000000000 0000000000 0010101010 0010101000 0010000001 0000010000 0000000000 0000010000 0000000000 0100000000 0000000   [000] 000 0000010000 1000000001 0000001000 0110000001 0000000000 0000010001 0000000000 0010000010 001
Gnetum        [??] 1000000000 0000000000 0000101010 0000100000 0010000000 0100000000 0000000010 0000011000 0000100000 0100000000 0000000   [000] 000 0000110010 0000000001 0000001000 0100000001 0000000000 0000010001 0000000010 0000001000 ???
Chloranthus   [??] 1000000000 0010100000 1000000000 0011110000 0000001000 0101010000 1000000000 0000010000 0000000000 0000000000 ???????   [000] 010 0000100000 1001000000 0100000000 0000000001 0000000000 0000000000 0000001000 0010000000 110
Piper         [??] 1001000000 0010001100 1000000000 0011110100 0000000000 0000010000 1000000000 1000000001 0000000000 0000000001 0000010   [?00] 000 0000100100 1001000000 0000000000 0100000000 0000000000 0010000000 0000001000 0010000000 010
Drimys        [??] 1000000001 0010000000 1000010000 0011110000 0000001000 0001010000 1000000000 0000010000 0000000000 0000000000 0000010   [000] 000 0000100000 1001000000 0100000000 0000100001 0100000000 1000010100 0100001000 0010000000 110

Calycanthus   [??] 1000000001 0010000000 0000000000 0011110000 0000001000 0100010000 1000000000 0000010000 0000000000 0000000000 0000010   [?00] 000 0000100000 1001000000 0100000000 1001000001 0000000000 0000000100 0100001000 0010000000 110
Eupomatia     [??] 1000000001 0010000000 0000000000 0001110001 0000001000 0101010000 1000000000 0010010010 0000000000 0000000000 0000010   [?00] 000 0000100000 1001000001 0100000000 0000000011 0100000100 0000000100 0100011000 0010000000 110
Magnolia      [00] 1000000001 0010100000 0000000000 0011110000 0000001000 0101010000 1000000000 0010010010 0000000000 0010000000 0000010   [000] 000 0000100000 1001000000 0100000000 1000000001 0100000000 1000010000 0000011000 0010000000 110
Persea        [00] 1000000001 0010100000 0000000000 0011110000 0000001000 0100010000 0000000000 0000010000 0000010000 0000000010 0000000   [000] 000 0000100000 1001000001 0100000000 0001000001 0000000000 0000000100 0100001000 0011000000 000
Trochodendron [??] 1000000001 1010100000 0000000000 0011100100 0001001000 0101010000 0000000000 0000110000 0000000001 0100000100 0000000   [?00] 000 0010100000 1000000000 0100000000 0000100001 0000000000 0000000100 0000101000 0010001000 010

Ceratophyllum [??] 0000000001 0010101000 0000100000 0011110001 0000001000 0001010000 0000000000 0001010000 0000100000 0000000000 0000000   [000] 000 0000100000 1000000001 0000000000 0000100001 0100000000 0000010000 0000001000 0010000000 010
Nymphaea      [??] 1001000001 0010000000 0000010000 0011110000 0000001000 0000010000 0000000000 0000010000 00000????? ?????????? ???????   [000] 000 0100100000 1001000000 0100000000 0100000001 0010000000 0000000000 0000001001 00?0000000 ???
Lilium        [??] 1000000001 0010000000 0000000000 0011110000 0000001000 0000010000 1000000001 0000010000 0000100000 0100000000 0100000   [?00] 000 0000000000 1001000000 0100000000 0100000001 0100000000 0000010000 0100001000 0010000000 000
Platanus      [??] 1000000001 0010100000 0000000000 0011100000 0000001000 0001010000 0000000000 0000010000 0000000000 0100000000 0000000   [?00] 000 0000100000 1000000000 0100000000 0000100000 1000000000 0000000100 0100101000 0010000000 100
Caltha        [??] 1000000011 0010000100 0000000000 0011100001 0000001000 0001000000 0000000000 0000010000 0001000001 0100000100 0000000   [000] 100 0000100000 1000000000 0100000000 0000001001 0000000100 0000000100 0000101000 0010000000 000

Dillenia      [??] 1000000001 0010000100 0000000000 0011100000 0000001000 0101010000 0000000000 0000010000 0000000000 0100000000 1000000   [?00] 000 0000100000 1000000000 0100000000 0000011001 0100000000 0000000100 0100000000 0010000000 000
Chrysolepis   [??] 1000000001 0011000000 0010000000 0001100000 0000001000 0101010000 0001000000 0000000000 0000000000 0100000101 0000000   [?00] 000 0000100000 1000000000 0100001010 0000010001 0000000000 0000000000 0000000000 0010000000 000
Betula        [??] 1000000001 0010000000 0000000000 0011100000 0010001000 0101010000 0000000000 0000000000 0000000000 0100000101 0000000   [000] 000 0000100000 1000000000 0100001010 0000010001 0100000000 0000000100 0100100000 0010000000 000
Casuarina     [??] 1000000001 0010110000 0000000000 0011100000 0001001000 0101010000 0000000000 0000000000 0000100000 0100000101 0000000   [?00] 000 0000100000 1000000000 0100001000 0100000001 0100000100 0000000100 0100100000 0010000000 000
Hamamelis     [??] 1000000001 0010000100 0000000000 0011100000 0001001000 0101010000 0000000000 0000010000 0000000001 0100000100 0000000   [?00] 000 0000100000 1000000000 0100000000 0100000001 0010000000 0000000100 0100001000 0010000000 000
```